2011

**Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression**

Lukas D. Wartman  
*Washington University School of Medicine in St. Louis*

David E. Larson  
*Washington University School of Medicine in St. Louis*

Zhifu Xiang  
*Washington University School of Medicine in St. Louis*

Li Ding  
*Washington University School of Medicine in St. Louis*

Ken Chen  
*Washington University School of Medicine in St. Louis*

*See next page for additional authors*

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

**Recommended Citation**

[https://digitalcommons.wustl.edu/open_access_pubs/1671](https://digitalcommons.wustl.edu/open_access_pubs/1671)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors
Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression

Lukas D. Wartman,1 David E. Larson,2 Zhifu Xiang,1 Li Ding,2 Ken Chen,2 Ling Lin,2 Patrick Cahan,1 Jeffery M. Klco,3 John S. Welch,1 Cheng Li,1 Jacqueline E. Payton,3 Geoffrey L. Uy,1 Nobish Varghese,2 Rhonda E. Ries,1 Mieke Hoock,1 Daniel C. Koboldt,2 Michael D. McAllan,2 Heather Schmidt,2 Robert S. Fulton,2 Rachel M. Abbott,2 Lisa Cook,2 Sean D. McGrath,2 Xian Fan,2 Adam F. Dukes,2 Tammi Vickery,2 Joelle Kalicki,2 Tamara L. Lamprecht,1 Timothy A. Graubert,1 Michael H. Tomasson,1 Elaine R. Mardis,2 Richard K. Wilson,2 and Timothy J. Ley1

1Departments of Internal Medicine and Genetics, Division of Oncology, Stem Cell Biology Section, Washington University School of Medicine, Siteman Cancer Center, St. Louis, Missouri, USA. 2The Genome Institute at Washington University, Washington University School of Medicine, St. Louis, Missouri, USA. 3Department of Pathology and Immunology, Washington University, St. Louis, Missouri, USA.

Acute promyelocytic leukemia (APL) is a subtype of acute myeloid leukemia (AML). It is characterized by the t(15;17)(q22;q11.2) chromosomal translocation that creates the promyelocytic leukemia–retinoic acid receptor α (PML-RARA) fusion oncogene. Although this fusion oncogene is known to initiate APL in mice, other cooperating mutations, as yet ill defined, are important for disease pathogenesis. To identify these, we used a mouse model of APL, whereby PML-RARA expressed in myeloid cells leads to a myeloproliferative disease that ultimately evolves into APL. Sequencing of a mouse APL genome revealed 3 somatic, nonsynonymous mutations relevant to APL pathogenesis, of which 1 (Jak1 V657F) was found to be recurrent in other affected mice. This mutation was identical to the JAK1 V658F mutation previously found in human APL and acute lymphoblastic leukemia samples. Further analysis showed that JAK1 V658F cooperated in vivo with PML-RARA, causing a rapidly fatal leukemia in mice. We also discovered a somatic 150-kb deletion involving the lysine (K)-specific demethylase 6A (Kdm6a, also known as Utx) gene, in the mouse APL genome. Similar deletions were observed in 3 out of 14 additional mouse APL samples and 1 out of 150 human AML samples. In conclusion, whole genome sequencing of mouse cancer genomes can provide an unbiased and comprehensive approach for discovering functionally relevant mutations that are also present in human leukemias.

Introduction

Recent studies have shown that whole genome sequencing is a valid, unbiased approach that can identify novel mutations and structural variants that may be important drivers of acute myeloid leukemia (AML) (1–3). However, human genomes have 3–4 million sequence variants per individual, and because cancers, such as AML, arise from individual HSCs that probably contain a large number of background mutations, it remains a difficult task to identify the somatic variants that are relevant for disease pathogenesis (1, 2). This task could be simplified considerably by sequencing the genomes of tumors arising in inbred strains of mouse cancer models, since the coverage requirements for detecting somatic mutations are predicted to be considerably less than those for sequencing outbred genomes, and the sequencing of a paired normal sample should not be absolutely required. In this study, we used a mouse model of APL to test this idea.

Acute promyelocytic leukemia (APL) (FAB M3 AML) is a subtype of AML characterized by the t(15;17)(q22;q11.2) translocation that creates the fusion oncogene, promyelocytic leukemia–retinoic acid receptor α (PML-RARA). Our laboratory and others have previously modeled APL in the mouse, in an effort to understand the genetic events that lead to leukemia (4–7). In our knockin mouse model, a human PML-RARA cDNA was targeted to the S′ untranslated region of the mouse cathepsin G (Ctsg) gene on chromosome 14 (referred to as mCG-PR mice) in an embryonic stem cell line that was derived from 129/SvJ mice (6). PML-RARA is expressed in the myeloid progenitor cells of these mice, and they develop a myeloproliferative disease that evolves into acute leukemia, with promyelocytic features after a latent period of 7–18 months. In the present study, F1 129/SvJ × C57BL/6 (B6) mCG-PR mice were backcrossed onto the B6/Taconic background (B6/T) for 10 generations, before establishing a large cohort of at-risk mCG-PR mice with sample banking. About 60% of the mCG-PR mice in the B6/T develop a disease that closely resembles APL after a long latent period, suggesting that additional progression mutations are required for APL development. Recurring, acquired chromosomal aberrations and distinct gene expression signatures are present in this and other mouse models of APL (8–10), but the spectrum of mutations required for progression are not yet known. To identify these mutations, we sequenced a diploid (as defined by a whole genome comparative genomic hybridization (CGH) tiling array; Supplemental Figure 1; supplemental material available online with this article; doi:10.1172/JCI45284DS1) mouse APL genome, using massively parallel DNA sequencing on the Illumina platform.
Results

Next-generation sequencing and identification of point mutations in a mouse APL genome. The sequenced mouse APL genome (tumor 9500) was derived from the leukemic spleen of a male mouse (mouse 9500) that developed leukemia after 335 days; AML in this mouse was characterized by the typical leukocytosis (>100 K/μl), splenomegaly (890 mg), and aberrant coexpression of the cell surface markers Gr-1 and c-kit on the splenic tumor cells (splenocytes were 49% Gr-1+/c-kit+). Residual 129/SvJ genome segments were predicted to account for a very small fraction of the tumor genome, based on a comparison of tumor copy number variants (defined by CGH data) to defined 129/SvJ copy number variants (Supplemental Figure 2 and ref. 11).

We created 2 Illumina paired-end libraries from tumor DNA (insert sizes of 300–350 bp and 550–600 bp) and generated 59.64 billion bp of sequence with 3 full Illumina GAIIx sequencing runs; the reads that were successfully mapped generated 15.564× haploid coverage. Surprisingly, when compared with the mouse B6/Jackson reference genome, the tumor 9500 sequence data predicted 87,778 heterozygous single nucleotide variants (SNVs) and 23,439 homozygous SNVs (Figure 1). We did not prospectively bank germline DNA from mouse 9500, since we did not anticipate significant genetic variation from the tumor sample compared with that of the mouse reference genome. Since normal mouse 9500 DNA was not available to definitively assign the somatic status of these SNVs, we sequenced the 129/SvJ genome on the Illumina platform, using pooled DNA derived from the spleens of 6 young WT male mice, to generate 28.9× haploid coverage (D.E. Larson, unpublished observations). 4,951,238 SNPs were present, compared with the reference B6/Jackson genome. Using this data, we determined that 0.96% of the tumor 9500 genome was derived from residual 129/SvJ sequence after 10 backcrosses to B6/T mice (Figure 2; see Methods for calculations). After filtering out the 129/SvJ SNPs, only 32,807 potentially somatic, heterozygous SNV s remained. 17,179 of the remaining sequence variants in the tumor genome occurred in contiguous blocks of sequence that we attributed to genetic drift that occurred between the 129/SvJ-derived ES cell line used to generate the mCG-PR mouse line (created in 1994) and the cells obtained from the control 129/SvJ WT mice that were used for sequencing the 129/SvJ genome (obtained in 2009 from The Jackson Laboratory). Some blocks of variants were also probably caused by genetic drift between the B6/T mice used for the backcrossing of the mCG-PR mice and the B6/Jackson mice used to sequence the reference genome.

To remove these background variants, we excluded all blocks of 4 or more SNVs (both heterozygous and homozygous) located within 40,000 bp of each other (Supplemental Table 1). 361 of the potential tier 1 somatic SNVs (tier 1 variants are in coding sequences,
as described in ref. 2) were found in these contiguous blocks of sequence and were not considered further (Supplemental Table 2).

The remaining 31 putative tier 1 SNVs were analyzed using Sanger sequencing of PCR products made from the original tumor DNA or from B6/T spleen DNA or 129/SvJ spleen DNA as controls. Both control DNA samples were pooled from spleen DNA derived from 6 healthy young male mice. Eight out of thirty-one calls were true positives (i.e., present only in the tumor genome). Two of these mutations (Dip2b, C to A, A633, and Gm6924, G to A, G121) were synonymous and were not pursued further. The final 6 nonsynonymous mutations were screened for recurrence in 89 additional mouse APL tumor samples (all derived from the same original mouse model). Mutations in the Jarid2 (L915I) and Capns2 (N149S) genes occurred only in the tumor derived from mouse 9500. Four out of six of the nonsynonymous mutations were found in additional samples; 3 of these mutations (in C6, Adad1, and Dbm1) were found together in a tumor from a littermate of mouse 9500 and in 2 other tumors from mice that were littermates of each other and bred from the same mCG-PR father as mouse 9500. The inheritance pattern therefore suggests that these mutations represent spontaneous germline mutations from a common ancestor of the proband and the other affected mice and that they are unlikely to be relevant for pathogenesis. The other recurring mutation was in the pseudokinase domain of Jak1 (V657F) and was identified in 6 other mice that were not closely related to the proband (7 out of 90 tumors = 8% of screened tumors) (Supplemental Figure 3A). We also sequenced the entire pseudokinase domain of Jak1 in these 89 mouse APL samples and found 1 additional sample had 2 different mutations in the Jak1 gene: G881E and N845S. Finally, we sequenced the entire Jak1 gene in 14 out of the 89 tumors (including the sample with 2 Jak1 mutations) and identified no additional mutations.

The Jak1 V657F mutation is orthologous to the V617F mutation in human JAK2 and is identical to a recently described activating JAK1 pseudokinase domain mutation (V658F) found in human APL and acute lymphoblastic leukemia samples (refs. 12–15 and Figure 3A). Previous work from our group demonstrated that other gain-of-function mutations in JAK1 (V623A and T478S) are relatively rare in human AML (16, 17).

Figure 2
Heat map of residual 129/SvJ SNPs by chromosomal location. A heat map was generated to show the number of SvJ/129 SNPs present in the tumor 9500 genome per 1-Mb window divided by the number total SvJ/129 SNPs present in the same 1-Mb window, aligned to the mouse chromosome ideogram. The location of cathepsin G (Ctsg) is highlighted by an arrow. The mouse 9500 breeding strategy placed selective pressure to retain the genetically modified Ctsg locus on chromosome 14 (the PML-RARA cDNA was knocked into the S′ untranslated region of the Ctsg gene). The experimentally derived heat map shows a skewing of retained 129/SvJ SNPs surrounding the Ctsg locus, as expected, with a random distribution of 129/SvJ SNPs throughout the remaining chromosomes. Min., minimum.
Figure 3
Functional validation of the recurrent Jak1 mutation identified in the sequenced mouse APL genome. (A) The mouse Jak1 V657F (mJAK1) mutation is orthologous to the human JAK1 V658F (hJAK1) and human JAK2 V617F (hJAK2) mutations. B41, band 4.1 domain; SH2, src-homology domain; JH2, pseudokinase domain; JH1, kinase domain. (B) Mouse APL Jak1 V657F mutation allele frequency by 454 sequencing. The variant allele frequencies of the V657F mutation from tumor 9500 and the 6 other tumors with V657F mutations are shown. On the far right are the read counts for the controls of a WT 129/SvJ DNA pool (n = 6) and a WT B6/T DNA pool (n = 6). F, female; M, male. (C) Functional validation with human JAK1 cDNA MSCV retroviral constructs. IRES-GFP (WT/GFP) alone versus JAK1 WT-IRES-GFP (WT/JAK1 WT) versus JAK1 V658F (WT/JAK1 V658F) mutant-IRES-GFP retroviral constructs were transduced into WT B6/T or mCG-PR bone marrow from 6-week-old mice. The transduced marrow was transplanted into lethally irradiated WT B6/T mice. The JAK1 V658F mutant cooperates with PML-RARA (PR) in a mouse model of APL and causes a fatal leukemia; a Kaplan-Meier overall survival plot is shown. This result was replicated in 2 independent experiments. (D) Serial flow cytometric analysis of peripheral blood shows a massive expansion of GFP+/Gr-1+ cells in the mCG-PR/JAK1 V658F cohort compared with that in the other experimental cohorts and controls. Error bars are mean ± SEM. (E) Flow cytometric analysis of peripheral blood, bone marrow, and spleen at the end of 1 JAK1 functional validation experiment showed no significant expansion of GFP+/Gr-1+ cells in cohorts other than mCG-PR/JAK1 V658F mice. Each symbol represents an individual mouse. Error bars are mean ± SEM.
We defined the variant allele frequency of the Jak1 V657F mutation in the 7 tumors using deep read counts on the 454/FLX platform, as previously described (ref. 1 and Figure 3B). The variant allele frequency in tumor 9500 was 43.86%, suggesting that nearly all of the cells in that tumor contained the heterozygous mutation. This was confirmed by sequencing 48 individual tumor 9500 colonies (grown in methylcellulose to assure clonality) to determine the Jak1 V657F allele frequency: 44 colonies were heterozygous for the Jak1 V657F mutation, 1 colony was homozygous for the Jak1 V657F mutation, and 3 colonies were Jak1 WT (data not shown). We also determined that the mutant Jak1 V657F allele was expressed in tumors 9500 and 12 at 36.67% and 48.78% of the cDNA reads, respectively, using reverse-transcriptase PCR, followed by sequencing of the complementary DNA on the 454/FLX platform (data not shown). The phenotype of the 2 Jak1-mutated mouse APL tumors on the B6/T (tumors 9500 and 26) was not different from 13 other well-characterized B6/T tumors (Supplemental Figure 3B). Similarly, the 2 Jak1-mutated tumors did not have a distinct gene expression signature in comparison with that of the 13 tumors without the mutation. An unsupervised analysis of gene expression microarray data from the 15 mouse APL tumors revealed that the 2 Jak1-mutated tumors were not tightly clustered (Supplemental Figure 4). Using the significance analysis of microarrays (SAM) algorithm, we found no significantly dysregulated genes between tumors with or without the Jak1 mutation (data not shown).

We recognize that pathogenically relevant somatic mutations, in both coding and noncoding regions of the genome (and germline susceptibility loci), may have been missed using our approach. However, our studies do suggest that the total number of true mutations present in this genome is similar to or less than the total numbers
of mutations detected in sequenced human AML genomes. We parsed tumor 9500’s heterozygous, nonblock-associated mutations by repetitive content (as defined by the mm9 RepeatMasker track from the University of California Santa Cruz Genome Browser). This breakdown yielded 31 tier 1 (coding sequence) variants; 1,311 nonrepetitive, non–tier 1 mutations; and 5,101 repetitive, non–tier 1 mutations. If the somatic validation rate of 6 out of 31 was constant across tiers, then we expected about 250 of the nonrepetitive variants to be true positives. Thus, the number of nonrepetitive variants in the mouse tumor genome was actually smaller than the number of variants in most of our validated human AML genomes. Those range from approximately 200 to 1,000 variants per genome from the 25 validated human AML cases studied so far (D.E. Larson and T.J. Ley, unpublished observations). The results from the CGH study, shown as a heat map in Supplemental Figure 1, also revealed that tumor 9500 has a diploid genome, with no copy number variants detected by dense whole genome CGH array. This provides orthogonal validation of our sequencing results and further supports the idea that this genome is not “unstable.”

Functional validation of a recurrent Jak1 mutation identified in the sequenced mouse APL genome. To test the functional significance of the Jak1 V657F (or human V658F) mutation for leukemia development, we expressed human JAK1 cDNA (using MSCV-based retroviral transduction) in bone marrow cells derived from healthy 6-week-old WT or mCG-PR mice in 2 independent experiments. Nine out of ten mice transplanted with mCG-PR bone marrow expressing the JAK1 WT or GFP constructs did not develop disease in this time frame (Figure 3C). The phenotype of the disease included splenomegaly (mean spleen weight of 449 ± 75 mg, n = 8) and leukocytosis (mean wbc count of 17.3 ± 12.0, n = 4). The mutant JAK1 V658F virus did not cause a detectable phenotype when transduced into WT marrow cells (Figure 3E). Immunoblots demonstrated that the levels of JAK1 protein were similar in splenocytes from the JAK1 WT and JAK1 V658F cohorts (data not shown).

The leukemia that developed in the JAK1 V658F mCG-PR animals was transplantable to healthy secondary recipients with a short latency (30–60 days). Southern blot analysis of spleen DNA from 5 of the diseased mice in the primary cohort revealed a small number of clonal integration sites in 4 tumors (oligoclonal) and no detectable clonal integration events in 1 tumor (Supplemental Figure 5). The myeloid cells of the primary JAK1 V658F mCG-PR tumors were more mature than those of the tumors that developed in the secondary recipients, for unclear reasons (Supplemental Figure 6).

A third retroviral transduction/transplantation experiment was performed using a lower dose of the MSCV vectors (~3% vs. 15% transduction efficiency in the prior experiments) to allow for quantitative assessment of myeloid expansion; peripheral blood samples from all mice were analyzed weekly. The mice from all cohorts in this experiment were sacrificed simultaneously at 10 weeks after transduction, when the mCG-PR/JAK1 V658F mice became ill. We detected a marked expansion of Gr-1+/GFP+ cells in the peripheral
blood of the mice transplanted with mCG-PR marrow transduced with the JAK1 v658F construct (Figure 3D). By 10 weeks, these mice had also developed splenomegaly and leukocytosis (Supplemental Figure 7). None of the other cohorts of mice exhibited this phenotype; however, there was a trend for expansion of Gr-1+/GFP cells in the cohort of mice with the JAK1 WT virus transduced into mCG-PR marrow (Figure 3E); the JAK1 WT mice did not develop myeloproliferation or early leukemias, demonstrating the importance of the V658F mutation for these phenotypes.

We tested the efficacy of the pan-JAK inhibitor (pyridone 6 [P6]) versus a selective JAK3 inhibitor (WHI-P131) in 10 banked mCG-PR tumors, using a methylcellulose colony-forming assay (Supplemental Figure 8). Banked mCG-PR tumors were thawed, incubated in liquid culture with or without drug for 48 hours, and then plated in methylcellulose for 5 to 7 days before colonies were counted. Although there was some variability in drug sensitivity among the 10 tumors, the pan-JAK inhibitor exhibited similar efficacy as that of ATRA in this assay (P = 0.6890). There was no significant difference in the efficacy of P6 between the tumors with the jaki v657 mutations (tumors 9500 and 26) versus those without jaki mutations (P = 0.0674). Using identical culture conditions, we showed that P6 hits its “target” in the tumor cells, because it decreased Stat5 phosphorylation (compared with DMSO-treated or JAK3 inhibitor-treated controls) in 2 mCG-PR tumors without jaki mutations (Supplemental Figure 9).

**Structural variant analysis of the sequenced mouse APL genome.** We next performed structural variant analysis of the sequenced tumor genome and identified 482 putative intrachromosomal rearrangements (deletions >30 bp, inversions, small insertions, and tandem duplications) that were not present in the B6/Jackson reference genome (18). An assembly score based on the assembly quality (length of contig and number of reads supporting the contig) was assigned, and 8 high-confidence, genic deletions were identified. PCR validation of 7 out of 8 of these regions showed evidence of a copy number variation (CNV) in 129/SvJ DNA, demonstrating that these were inherited variants from the 129/SvJ strain (data not shown). However, we validated a somatic 150-kb deletion on chromosome X that included nearly all of the lysine (K)-specific demethylase 6A (Kdm6a, also known as Utx) gene, a histone H3K27 demethylase that has recently been shown to be deleted in 2 human AML cell lines and several other cancers (refs. 19, 20, and Figure 4, A and B). We screened 14 additional mouse APL tumors using an ultra-dense Nimblegen CGH custom-tiling array and found Kdm6a deletions in 4 out of 15 tumors, including the proband (Figures 5 and 6 and Table 1). The expression of Kdm6a was reduced to background levels in the tumors containing the deletions (Supplemental Figure 10A). We screened 150 human AML samples for deletions in the KDM6A region of the X chromosome, using paired tumor and skin DNA samples on the Affymetrix v6.0 SNP array platform. One female patient with FAB M2 AML, whose cytogenetics and FISH analysis revealed t(8;21) and loss of X, had a ~200-Kb deletion of the remaining X chromosome (which included the DUSP21 gene and the 5′ region and first 3 exons of KDM6A) (Figure 4C, Figure 6, and Supplemental Figure 11). Although KDM6A is normally expressed in CD34 cells and in nearly all AML samples, the expression of KDM6A was undetectable in this AML case, validating the transcriptional consequences of the deletion plus the loss of 1 X chromosome (Supplemental Figure 10B). The role of this mutation in AML pathogenesis is not yet clear but may involve the dysregulation of Hox, Notch, or Rb signaling pathways (21–24). The 4 Kdm6a-deleted mouse APL tumors did not have a distinct gene expression signature in comparison with that of 11 tumors without the deletion (using unsupervised clustering analysis or a supervised analysis comparing expression of only the Hox genes; Supplemental Figures 4 and 12). The only significantly dysregulated gene between tumors with and without Kdm6a deletions was Kdm6a itself (data not shown). These data suggest that the 11 tumors without Kdm6a deletions may also have mutations in this gene or mutations in the pathway that this gene affects.

**Discussion**

Recurring, acquired chromosomal aberrations and distinct gene expression signatures have been previously identified in mouse models of APL (8–10), but prior studies, using array-based platforms, have not been able to identify the relevant driver mutations that arise during disease progression. By sequencing a single mouse...
A.P.L. genome using next-generation technologies, we have identified a recurring point mutation that is relevant for A.P.L. pathogenesis (*Jak1 V657F* or V658F in humans) and a recurring deletion in a histone demethylase (*Kdm6a*) that is present in both human and mouse A.M.L. genomes. *Kdm6a* and other genes involved in histone methylation, such as *JARID1C* and *EZH2*, have recently been found to be mutated in a variety of human cancers, including myeloid malignancies (19, 20, 25–28). We have shown that aberrant *Jak1* signaling can play an important role in A.P.L. pathogenesis in vivo; therefore, A.P.L. samples with or without *Jak1* mutations are sensitive to a pan-JAK inhibitor in a colony-forming assay, which suggests that JAK-STAT signaling may be activated by additional mechanisms in this mouse model of A.P.L.

Finally, by sequencing a mouse cancer genome derived from a mouse backcrossed 10 generations on the B6/T, we have learned that a significant amount of genetic heterogeneity is present even in inbred mouse strains; therefore, a matched normal sample will be optimal for defining the somatic status of mutations arising in mouse models of cancer. Regardless, the approach described here may allow us to rapidly identify progression mutations that are relevant for human A.M.L. and provides an important proof of concept that this mouse model of A.M.L. is functionally related to its human counterpart. Mutations in several known oncogenes, including *K-Ras* and *Flt3* (and dysregulated expression of *Bcl2*), have been previously shown to cooperate with *Pml-Rara* in this mouse APL model (29–33); importantly, these mutations occur in many A.M.L. subtypes, not just A.P.L. The mutations that spontaneously arise in this mouse model of human A.P.L. can now be identified by whole genome sequencing using next-generation platforms; this provides a physiologic system for the detection of leukemia progression mutations that are also relevant for human A.M.L. pathogenesis.

### Methods

**Preliminary screen of somatic copy number alterations.** DNA from 15 mouse A.P.L. tumors on the B6/T (F10) and pooled DNA derived from spleen of 6-week-old WT, parental strain mice (B6/T, n = 6, and 129X1/SvJ, n = 6) was hybridized to the NimbleGen Mouse CGH 3x720K WG-T arrays, as previously described (34). To systematically identify gross chromosomal gains and losses, we applied the wuHMM algorithm to the average log2 ratios of 50 consecutive probes (35). Because segmentation algorithms assume that the test and reference DNA samples have largely the same copy number across a chromosome, gross abnormalities can confound the identification of micro-alterations. To compensate for this, we mean centered each chromosome, except for chromosomes containing gross abnormalities. For each of these chromosomes, we mean centered the called and non-called regions separately. To identify micro-alterations, we executed wuHMM on the normalized individual probe-level data, identifying 9,897 putative CNVs. Next, we determined the extent of residual 129/SvJ CNVs in each tumor. First, we identified 140 germline CNVs in 129/SvJ versus B6/T (60 of the 140 CNVs were identified as amplifications with a gain score of >1.5, and the remaining 80 CNVs were identified as deletions with a loss score of >2.5; gain and loss scores were calculated using the wuHMM algorithm, in which CNVs are scored based on log2 ratio magnitude, number of probes, and local noise). For each CNV region, we assigned genotypes to tumor samples based on the log2 ratio of individual probes within the region, as previously described (36). CGH array data for all samples have been deposited on the Gene Expression Omnibus (GEO; [http://www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/); SuperSeries accession no. GSE26131).

**Tiling-array validation of somatic copy number alterations in 15 tumors.** To prioritize candidate copy number alterations (CNAs) for validation on a custom tiling array, first we filtered calls based on visual inspection. This process resulted in 455 calls. 84 out of the 455 CNAs substantially overmapped 129X1/SvJ CNVs that we identified in this study. These were considered germline polymorphisms and were excluded from further analyses, leaving 371 candidates. It is possible that we missed real CNAs in the visualization process. Therefore, of the calls that did not pass our initial inspection, we also included 268 calls that passed stringent score thresholds: gains scoring of more than 1.25 and losses scoring of more than 1.75. Overlapping CNAs were merged into nonredundant regions, leading to a final list of 449 candidate regions for validation by custom-tiling array. Validation arrays were designed such that each region was covered by approximately 300 probes.

**Mouse gene expression arrays.** Total cellular RNA was purified using TRIzol reagent (Invitrogen), quantified using UV spectrophotometry (Nanodrop Technologies), and qualitatively assessed using an Experion Bioanalyzer. Amplified cDNA was prepared from 20 ng total RNA using the whole transcript WT-Ovation RNA Amplification System and biotin-labeled using the Encore Biotin Module, both from NuGen Technologies, according to the manufacturer’s instructions. Labeled targets were then hybridized to Mouse Exon 1.0 ST arrays (Affymetrix), washed, stained, and scanned using standard protocols from the Siteman Cancer Center, Molecular and Genomic Analysis Core Facility ([http://pathology.wustl.edu/research/cores/lcg/index.php](http://pathology.wustl.edu/research/cores/lcg/index.php)). Affymetrix Expression Console software was used to process array images, extract signal data, and evaluate image and data quality relative to standard Affymetrix quality control metrics. The expression data from 3 out of the 34 probe sets for *Kdm6a* was plotted; these 3 “best” probe sets had the highest expression with the lowest variability in the tumors without the *Kdm6a* deletion. The average expression of *Kdm6a* in the 11 tumors without *Kdm6a* deletions was significantly higher than the expression of the tumors with deletions involving *Kdm6a*, if the raw expression data from all 34 probe sets was averaged for each sample (*P* < 0.0001). Both the unsupervised and supervised heat map analysis of the 4 *Kdm6a*-deleted
tumors versus the remaining 11 tumors was done using Spotfire software (TIBCO). To determine significantly dysregulated genes, we removed probes with raw signal intensity of less than 200 in all samples (bringing the total number of probes in the analysis down to 142,000) and used a fold-change cutoff of less than or equal to 0.5 or more than or equal to 2 and P-value cutoff of less than or equal to 0.05, using the ANOVA algorithm. Finally, we removed probe sets with average expression of less than 200 in both groups. The remaining probe sets were also analyzed using SAM (37); 6 probe sets were significant at an FDR of 0, and all 6 probe sets were significant for Kdm6a. We used the same approach for the analysis of the 2 Jak1-mutated tumors in comparison with the 13 tumors without a Jak1 mutation. Exon array data for all samples have been deposited on GEO (http://www.ncbi.nlm.nih.gov/geo/; SuperSeries accession no. GSE26131).

**Human gene expression and SNP arrays.** The gene expression array and analysis of the human sample 750152 was done on the Affymetrix Human v1.33 Plus 2 array, as previously described (38). The copy number analysis of the human sample 750152 was done by Affymetrix v6.0 SNP array, as previously described (39), and analyzed by the CMDS algorithm (40). One hundred fifty human AML samples were screened by CMDS of Affymetrix v6.0 SNP array data, with a focal deletion involving the KDM6A gene in 1 sample, 750152. This study was approved by the Human Research Protection Office at Washington University School of Medicine, after patients and donors provided informed consent, in accordance with the Declaration of Helsinki.

**Illumina library construction.** We followed the standard library construction procedure, as previously described (16). Briefly, we started with 500 ng DNA from the mouse APL tumor 9500 and 1 μg DNA from a pool of equal amounts of DNA isolated from the spleens of 6 young 129X1/SvJ mice. After the DNA was fragmented, end-repaired, and ligated with adaptors, we amplified the products and ran them on a polyacrylamide gel. We then excised gel slices between 350 and 400 bp and 550 and 600 bp for the tumor 9500 sample and 300 and 350 bp for the 129X1/SvJ pool and eluted the DNA from each gel slice as the source for independent libraries.

**Illumina sequencing.** After diluting the libraries to a 10-nm concentration, we used the paired-end flow cell and cluster generation kits to produce flow cells with an average cluster density ranging between 115,000 to 420,000 clusters per tile. We used the standard sequencing kits (Illumina Sequencing Kit v3) and performed 36–100 cycles of nucleotide incorporation. After the DNA was fragmented, end-repaired, and ligated with adaptors, we amplified the products and ran them on a polyacrylamide gel. We then excised gel slices between 350 and 400 bp and 550 and 600 bp for the tumor 9500 sample and 300 and 350 bp for the 129X1/SvJ pool and eluted the DNA from each gel slice as the source for independent libraries.

**Alignment, coverage analysis, and SNV/indel calling.** Illumina reads from both mouse strains were aligned to the mouse NCBI build 37 reference sequence using MAQ (41). Haploid coverage for each genome was estimated using MAQ’s mapcheck command at 28.9X for 129/SvJ and 15.5X for tumor 9500. SNVs in tumor 9500 were detected by MAQ and further filtered using MAQ SNPFilter. Despite 10 generations of backcrossing to B6/T, we expected to find a small number of 129/SvJ variants clustered in inherited blocks. Thus, we predicted variants from our genomic reads of 129/SvJ with SAMtools and removed tumor 9500 variants sharing an allele with predictions from 129/SvJ from consideration (42). We then further removed predicted sites occurring in blocks. We defined a block as a cluster or at least 4 variants of which each variant was less than 40,000 nucleotides away from its nearest neighbor. After filtering, 31 potential tier 1 heterozygous SNVs in noncontiguous blocks were then validated by resequencing on the 3730 DNA Analyzer platform, and 23 out of 31 sites did not validate. Of these, 1 had no coverage, and 3 had low-quality data due to simple repeat content at the site. Six were variants in our B6/T WT pool and thus represented variation from the sequenced reference in our backcross. One site was present in both the B6/T WT pool and the 129/SvJ WT control, but the 129/SvJ variant did not pass MAQ SNPFilter and hence was not identified as a 129/SvJ variant. One variant was present in the 129/SvJ sequence but did not pass MAQ SNPFilter and thus was not filtered out by our analysis. Eleven sites validated as matching the reference sequence. Of those, manual review found that 2 had low coverage, 3 looked like potential paralog mappings, 2 were near homopolymers, 2 were in regions with reads supporting the variant that were ambiguously mapped (in addition to some reads that were unambiguous), and 2 gave no indication of why they might have validated WT.

**Determination of residual 129/SvJ genomic regions.** We found 4,951,238 SNPs present in the 129/SvJ genome compared with the B6/Jackson reference genome. There were 78,339 129/SvJ SNPs present in the tumor 9500 genome (10,479 were homozygous; 55,691 were heterozygous but overlapped homozygous 129/SvJ mutations; and 12,169 were heterozygous and exactly matched heterozygous mutations in 129/SvJ). The percentage of B6/Jackson genome in the tumor 9500 genome equals ([0.5 × number of heterozygous SNPs in tumor 9500 that were exactly matched heterozygous SNPs in 129/SvJ genome] × [0.75 × number of heterozygous SNPs in tumor 9500 that exactly matched heterozygous SNPs in 129/SvJ genome] + [number of homozygous SNPs present in both tumor 9500 and 129/SvJ genome])/(total number of polymorphic alleles). Thus, the percentage of residual 129/SvJ genome in the tumor 9500 genome was calculated to be ([0.5 × 55,691 SNPs] + [0.75 × 12,169 SNPs] × 10,479 SNPs)/4,951,238 alleles or 9.6%. Conversely, the percentage of the B6/Jackson genome present in the tumor 9500 genome was 99.04%. We used the Circos program (http://mkweb.bcgsc.ca/circos/) to generate the heat map of residual 129/SvJ SNPs by chromosomal location (Figure 2 and ref. 43).

**Determination of Jak1 V657F allele frequency.** We thawed a vial of passaged tumor 9500 and then plated the tumor cells using serial dilution in methylcellulose plates containing SCF, IL-3, and IL-6 (MethoCult M3534, Stem Cell Technologies) for 7 days before 48 distinct, individual colonies (>50 cells) were harvested by direct visualization under light microscope. Individual colonies were transferred to a single well of a 96-well plate, washed with PBS, and spun down. Cell pellets were then directly used as the input for whole genome DNA amplification, as described in ref. 2. Whole genome amplification DNA was then used as the input DNA for standard PCR amplification of a Jak1 amplicon, followed by Sanger sequencing. We included 2 mouse APL tumors with known Jak1 V657F mutations, 2 mouse APL tumors without Jak1 mutations, and 2 blank wells as controls for this experiment.

**Sequence based copy number analysis.** We aligned the APL data with BWA and counted the number of mapped reads in consecutive nonoverlapping 1-kb intervals across the genome (44). We estimated copy numbers by dividing the read counts by the chromosomal median. Copy number altered regions (CNARs) between these 2 strains were identified by first dividing the genome into 10-kb nonoverlapping windows. We estimated the copy number value of each window as twice (diploid) the ratio of the number of reads confidently mapped (mapping quality >35) within the window over the median of all the windows in the genome. For each genome, we used a Hidden Markov Model (HMM) to segment the copy number values and produce copy number segments. Each state in the HMM corresponds to a discrete copy number, and the distribution of the copy number values given a state was modeled as a Gaussian probability density function. We used the expectation maximization algorithm to estimate the HMM parameters and determined the segmentation using the Viterbi algorithm. We derived CNARs by comparing the segmentation of the tumor genome versus that
technical advance

of the normal genome. A log likelihood ratio (LLR) score was computed for each CNAR to quantify the possibility of being a true CNAR from being the null (copy number neutral) hypothesis. All CNARs with LLRs of more than or equal to 100 were reported. We have named the above method for detecting CNARs, CNVHM. The boundaries of the CNARs were further refined using supporting discordant read pairs identified by BreakDancer (18). PCR validation was done with primers designed upstream and downstream of the predicted deletion breakpoints. For tumor 9500, the forward primer was GACAGACCTACCTCTCATG at position 17,752,753 of chromosome X, and the reverse primer was GACAGACCTACCTCTCATG at position 17,904,755 (using build 37.1 of the mouse genome). The PCR product was cloned using the Topo Ta Cloning Kit (Invitrogen) and confirmed by full-length DNA sequencing. Retroviral supernatants were generated and viral titers determined by flow cytometry to quantify the possibility of being a true CNAR from being mutagenized using the QuikChange Site-Directed Mutagenesis Kit (Stratagene) and confirmed by full-length DNA sequencing. Retroviral expression vector (MSCV retroviral constructs) was obtained from a commercial vendor (Origene) and subcloned into the retroviral vector expression vector MSCV-IREs-EFP (MIG). The V658F mutation was introduced into the JAK1 coding sequence by site-directed mutagenesis using the QuickChange Site-Directed Mutagenesis Kit (Stratagene) and confirmed by full-length DNA sequencing. Retroviral supernatants were generated and viral titers determined by flow cytometry analysis of the GFP+ cells in retrovirally transfected 3T3 cells, as described previously (45). GFP alone versus JAK1 WT/IREs/EFP versus JAK1 V658F mutant/IREs/EFP MSCV retroviral constructs were transduced into WT B6/T or mCG-PR bone marrow from 6-week-old healthy mice. The transduced marrow was then transplanted into lethally irradiated (1,100 cGy) syngeneic B6/T mice. Secondary transplants were performed following standard protocols as described previously (6). The Jak1 (6G4) rabbit mAb used was from Cell Signaling Technology. To assess protein loading, blots were probed with anti-β-actin antibody (Sigma-Aldrich).

**DNA isolation and Southern analysis.** Genomic DNA was isolated from banked splenocytes after being lysed with a 1% Triton X-100 lysis buffer with 2 mg/ml Proteinase K added (Sigma-Aldrich), followed by standard phenol-chloroform extraction. For assay of clonality, 10 μg genomic DNA was digested with EcoRI or HindIII (as a positive control) (both from Fisher Scientific). The hybridization probe was a 277-bp PCR fragment generated from primers targeting the cDNA of the EGFP gene.

**Statistics.** Statistical comparisons were made using Student’s 2-sided t test unless otherwise noted. We performed the above calculations and the Kaplan-Meier survival analysis using the log-rank test with the GraphPad Prism 5 software. We considered P ≤ 0.05 to be significant, except for the analysis of the methylcellulose colony-forming assay (Supplemental Figure 8), for which we considered a cutoff of P ≤ 0.01 as significant.

**Acknowledgments**

This work was supported by NIH grants CA083962 and CA101937 and the Barnes-Jewish Hospital Foundation grant 00335-0505-02 to T.J. Ley. The Siteman Cancer Center and High Speed Cell Sorter Core is supported in part by an NCI Cancer Center Support Grant, P30 CA91842. The Microarray Core is supported by National Center for Research Resources grant UL1 RR024992. L.D. Wartman was supported by NIH grant T32 HL007088.

Received for publication September 30, 2010, and accepted in revised form January 19, 2011.

Address correspondence to: Richard K. Wilson, The Genome Institute at Washington University, 4444 Forest Park Blvd., Box 8501, St. Louis, Missouri 63108, USA. Phone: 314.286.1814; Fax: 314.286.1810; E-mail: rwilson@wustl.edu.

Zhifu Xiang’s present address is: Department of Medicine, Division of Oncology, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA.

Rhonda E. Ries’s present address is: Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.

Patrick Cahan’s present address is: Division of Hematology/Oncology, Children’s Hospital, Boston, Massachusetts, USA.


technical advance