ORIGINAL ARTICLE

Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome

Elaine R. Mardis, Ph.D., Li Ding, Ph.D., David J. Dooling, Ph.D., David E. Larson, Ph.D., Michael D. McLellan, B.S., Ken Chen, Ph.D., Daniel C. Koboldt, M.S., Robert S. Fulton, M.S., Kim D. Delehaunty, B.A., Sean D. McGrath, M.S., Lucinda A. Fulton, M.S., Devin P. Locke, Ph.D., Vincent J. Magrini, Ph.D., Rachel M. Abbott, B.S., Tammi L. Vickery, B.S., Jerry S. Reed, M.S., Jody S. Robinson, M.S., Todd Wylie, B.S., Scott M. Smith, Lynn Carmichael, B.S., James M. Eldred, Christopher C. Harris, B.S., Jason Walker, B.A., B.S., Joshua B. Peck, M.B.A., Feiyu Du, M.S., Adam F. Dukes, B.A., Gabriel E. Sanderson, B.S., Anthony M. Brummett, Eric Clark, Joshua F. McMichael, B.S., Rick J. Meyer, M.S., Jonathan K. Schindler, B.S., B.A., Craig S. Pohl, M.S., John W. Wallis, Ph.D., Xiaoqi Shi, M.S., Ling Lin, M.S., Heather Schmidt, B.S., Yuzhu Tang, M.D., Carrie Haipek, M.S., Madeline E. Wiechert, M.S., Jolynda V. Ivy, M.B.A., Joelle Kalicki, B.S., Glendoria Elliott, Rhonda E. Ries, M.A., Jacqueline E. Payton, M.D., Ph.D., Peter Westervelt, M.D., Ph.D., Michael H. Tomasson, M.D., Mark A. Watson, M.D., Ph.D., Jack Baty, B.A., Sharon Heath, William D. Shannon, Ph.D., Rakesh Nagarajan, M.D., Ph.D., Daniel C. Link, M.D., Matthew J. Walter, M.D., Timothy A. Graubert, M.D., John F. DiPersio, M.D., Ph.D., Richard K. Wilson, Ph.D., and Timothy J. Ley, M.D.

ABSTRACT

BACKGROUND

The full complement of DNA mutations that are responsible for the pathogenesis of acute myeloid leukemia (AML) is not yet known.

METHODS

We used massively parallel DNA sequencing to obtain a very high level of coverage (approximately 98%) of a primary, cytogenetically normal, de novo genome for AML with minimal maturation (AML-M1) and a matched normal skin genome.

RESULTS

We identified 12 acquired (somatic) mutations within the coding sequences of genes and 52 somatic point mutations in conserved or regulatory portions of the genome. All mutations appeared to be heterozygous and present in nearly all cells in the tumor sample. Four of the 64 mutations occurred in at least 1 additional AML sample in 188 samples that were tested. Mutations in *NRAS* and *NPM1* had been identified previously in patients with AML, but two other mutations had not been identified. One of these mutations, in the *IDH1* gene, was present in 15 of 187 additional AML genomes tested and was strongly associated with normal cytogenetic status; it was present in 13 of 80 cytogenetically normal samples (16%). The other was a nongenic mutation in a genomic region with regulatory potential and conservation in higher mammals; we detected it in one additional AML tumor. The AML genome that we sequenced contains approximately 750 point mutations, of which only a small fraction are likely to be relevant to pathogenesis.

CONCLUSIONS

By comparing the sequences of tumor and skin genomes of a patient with AML-M1, we have identified recurring mutations that may be relevant for pathogenesis.

From the Departments of Genetics (E.R.M., L.D., V.J.M., R.K.W., T.J.L.), Medicine (R.E.R., P.W., M.H.T., S.H., W.D.S., D.C.L., M.J.W., T.A.G., J.F.D., T.J.L.), and Pathology and Immunology (J.E.P., M.A.W., R.N.); the Genome Center (E.R.M., L.D., D.J.D., D.E.L., M.D.M., K.C., D.C.K., R.S.F., K.D.D., S.D.M., L.A.F., D.P.L., V.J.M., R.M.A., T.L.V., J.S. Reed, J.S. Robinson, T.W., S.M.S., L.C., J.M.E., C.C.H., J.W., J.B.P., F.D., A.F.D., G.E.S., A.M.B., E.C., J.F.M., R.J.M., J.K.S., C.S.P., J.W.W., X.S., L.L., H.S., Y.T., C.H., M.E.W., J.V.I., J.K., G.E., M.A.W., R.K.W., T.J.L.); Siteman Cancer Center (P.W., M.H.T., M.A.W., S.H., W.D.S., R.N., D.C.L., M.J.W., T.A.G., J.F.D., R.K.W., T.J.L.); and the Division of Biostatistics (J.B.) — all at Washington University, St. Louis. Address reprint requests to Dr. Ley at Washington University, 660 S. Euclid Ave., Campus Box 8007, St. Louis, MO 63110, or at timley@wustl.edu.

This article (10.1056/NEJMoa0903840) was published on August 5, 2009, at NEJM.org.

N Engl J Med 2009;361:1058-66.

Acute myeloid leukemia (AML) is a clonal hematopoietic disease caused by both inherited and acquired genetic alterations.[1-3] Current AML classification and prognostic systems incorporate genetic information but are limited to known abnormalities that have previously been identified with the use of cytogenetics, array comparative genomic hybridization (CGH), gene-expression profiling, and the resequencing of candidate genes (see the Glossary).

The karyotyping of AML cells remains the most powerful predictor of the outcome in patients with AML and is routinely used by clinicians.[4,5] As an adjunct to cytogenetic studies, small subcytogenetic amplifications and deletions can be identified with the use of genomic methods, such as single-nucleotide-polymorphism (SNP) array and array CGH platforms (see the Glossary). However, these techniques remain investigational, and studies[6-9] suggest that there are few recurrent acquired copy-number alterations in each AML genome. Gene-expression profiling has identified patients with known chromosomal lesions and genetic mutations and subgroups of patients with normal cytogenetic profiles who have variable clinical outcomes.[10,11] Expression profiling has yielded single-gene predictors of outcome that are currently being evaluated for clinical use.[12-16] Candidate-gene resequencing studies have also identified recurrent mutations in several genes — for example, genes encoding FMS-related tyrosine kinase 3 (*FLT3*) and nucleophosmin 1 (*NPM1*) — that can help to stratify patients with normal cytogenetic profiles according to risk and to identify patients for targeted therapy (e.g., those with mutated *FLT3*).[3,12,17] However, the revised classification systems are imperfect, suggesting that important genetic factors for the pathogenesis of AML remain to be discovered.

We have previously described the sequence of an entire AML genome from a patient who had AML with minimal maturation (AML-M1) and a normal cytogenetic profile.[18] Here we describe the genome sequence of another such tumor and recurring mutations in additional AML tumors.

METHODS

Details regarding the methods for library production, DNA sequencing with the Illumina Genome Analyzer II,[19] evaluation of sequence coverage, identification of sequence variants, validation of variants and determination of the prevalence of variants in the index AML tumor, and screening of additional AML samples are provided in the Supplementary Appendix, available with the full text of this article at NEJM.org. All the high-quality single-nucleotide variants (SNVs) that were found in tumor and skin samples from this patient are available in the database of genotypes and phenotypes (dbGaP) of the National Center for Biotechnology Information (accession number, phs000159.v1.p1).

RESULTS

CASE REPORT

A previously healthy 38-year-old man of European ancestry presented with fatigue and a cough. The white-cell count was 39,800 cells per cubic millimeter, with 97% blasts; the hemoglobin level was 8.9 g per deciliter, and the platelet count was 35,000 per cubic millimeter. A bone marrow examination revealed 90% cellularity and 86% myeloperoxidase-positive blasts (Fig. 1 in the Supplementary Appendix). Routine cytogenetic analysis of bone marrow samples revealed a normal 46,XY karyotype. There was no family history of leukemia. The patient's mother had received the diagnosis of breast cancer at the age of 60 years and of non-Hodgkin's lymphoma at the age of 63 years; her half-sister had received the diagnosis of breast cancer at the age of 50 years.

Samples of the patient's bone marrow and skin were banked for whole-genome sequencing under a protocol approved by the institutional review board at Washington University. The patient provided written informed consent.

The patient was treated initially with a 7-day course of infusional cytarabine and with a 3-day course of daunorubicin. Within 5 weeks, he had complete morphologic remission and recovery of white-cell and platelet counts. The patient subsequently received consolidation therapy with four cycles of high-dose cytarabine without any further antileukemic therapy. He remained in complete remission 3 years later.

CHARACTERIZATION OF THE TUMOR GENOME

DNA samples from the patient's bone marrow sample at the time of initial presentation and a normal skin-biopsy specimen obtained after the patient's disease was in remission were labeled

Glossary

Build 36 of the human reference genome: The most current version of the assembled human genome reference sequence, available online at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/).

Comparative genomic hybridization (CGH): A comparison of DNA abundance, throughout the genome, between two DNA samples to identify regions where DNA copies have been gained or lost.

dbSNP: A publicly available database of known DNA variants, housed at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/SNP).

Diploid coverage: A metric used in whole-genome sequencing studies to describe the likelihood of detecting both alleles at any given nucleotide position in a genome.

Genic: Regions of the genome that contain genes, including their exons and introns.

Haploid coverage: A metric used in whole-genome sequencing studies to describe detection of each nucleotide position in a genome for at least one allele; "1× coverage" is equivalent to the size of the genome (e.g., approximately 3 billion base pairs for the human genome).

Next-generation sequencing: A variety of new techniques that have in common the generation of DNA sequence from single molecules of DNA, rather than pools of DNA templates; hundreds of millions of DNA fragments can be sequenced at the same time on a single platform (massively parallel sequencing).

Paired-end reads: DNA sequences that are produced on next-generation sequencing platforms by sequencing both ends of DNA fragments, resulting in higher confidence in assigning the sequence a position in the reference genome and allowing the detection of structural variations.

Partial uniparental disomy: An acquired somatic recombination event that causes the duplication of a part of a chromosome from one parent, resulting in a "copy-number neutral" loss of heterozygosity for a chromosomal segment.

Resequencing: Obtaining the DNA sequence of additional members of a species for which a completed reference sequence is known and to which comparisons can be made.

Sequencing run: The sequence that is generated by a complete Illumina flow cell or a similar next-generation sequencing platform; one sequencing run generates many billions of base pairs of sequence.

Single-nucleotide polymorphism (SNP): A position in the genome where some individuals in a population inherit a change in a single nucleotide that differs from the reference genome.

Single-nucleotide variant (SNV): A difference in a DNA sequence of a sample at a single position in the genome, as compared with the reference genome; each variant may represent either an inherited or an acquired change.

SNP array: A microarray-based assay system that allows for simultaneous measurement of nucleotide sequence and abundance in a DNA sample at possibly hundreds of thousands of positions in the genome.

and genotyped with the use of the Affymetrix Genome-Wide Human SNP Array 6.0. The tumor genome had no detectable somatic copy-number alterations and no regions of partial uniparental disomy (Glossary, and Fig. 2 in the Supplementary Appendix). RNA that was derived from the same bone marrow sample was analyzed with the use of the Affymetrix GeneChip Human Genome U133 Plus 2.0 array, which revealed an expression signature similar to that of many other cytogenetically normal marrow samples from patients with AML-M1 (Fig. 2 in the Supplementary Appendix).

SEQUENCE COVERAGE AND POTENTIAL MUTATIONS

We sequenced 69.9 billion base pairs (23.3× haploid coverage) from DNA libraries that we generated from the tumor sample and 63.9 billion base pairs from libraries that we generated from the normal skin sample (21.3× haploid coverage) (Glossary and Table 1). Using Affymetrix 6.0 SNP arrays, we confirmed the detection of both alleles of 98.5% of the approximately 45,000 high-quality heterozygous SNPs in the tumor sample and 97.4% of the approximately 45,000 high-quality heterozygous SNPs in the skin sample.

A summary of the sequence differences between the patient's tumor genome and National Center for Biotechnology Information build 36 of the human reference genome is shown in Figure 1 (see the Glossary).[20] We identified 3,872,936 SNVs in the tumor genome, of which 3,464,449 passed a stringent calling filter. Of these SNVs, 3,377,680 (97.5%) were detected in the skin genome, indicating that they were inherited variants. Of the 86,769 potentially novel somatic SNVs, 66,513 had been described previously.

We binned the remaining 20,256 SNVs into four tiers, which are detailed in the Supplementary Appendix. Briefly, tier 1 contains all changes in the amino acid coding regions of annotated exons, consensus splice-site regions, and RNA

genes (including microRNA genes). Tier 2 contains changes in highly conserved regions of the genome or regions that have regulatory potential. Tier 3 contains mutations in the nonrepetitive part of the genome that does not meet tier 2 criteria, and tier 4 contains mutations in the remainder of the genome. We tentatively identified 113 potential tier 1 mutations, 749 potential tier 2 mutations, 3188 potential tier 3 mutations, and 16,206 potential tier 4 mutations. For each of the 113 putative tier 1 variants, we amplified the genomic region containing the mutation from both tumor and skin, using a polymerase-chain-reaction (PCR) assay, and performed Sanger sequencing. Of the 101 variants that were called with low confidence (the calling algorithm is summarized in the Supplementary Appendix), none were validated. Of the high-confidence variants, 10 of 12 were validated as somatic mutations. Similarly, we tested 178 low-confidence calls for tier 2, and only one was validated. In contrast, 51 of 104 high-confidence tier 2 calls were validated. We did not carry out validation studies of variants in tiers 3 and 4.

We also searched for somatic insertions and deletions (indels) using an algorithm described in the Supplementary Appendix. We identified 142 potential somatic indels (28 deletions and 114 insertions). Of these variants, 119 failed validation (i.e., they were falsely positive) in Sanger sequencing of the relevant PCR products, 21 were validated but were present in both tumor and skin, and 2 were validated as somatic mutations. One was a 4-bp insertion in exon 12 of the *NPM1* gene associated with aberrant cytoplasmic expression of nucleophosmin (*NPM*c). This insertion creates a frameshift mutation and a truncated protein that is known to have altered cellular localization, as described previously.[21] The second mutation was a 3-bp insertion in the gene encoding centrosomal protein 170kDa (*CEP170*) at amino acid 177, predicted to result in the addition of a leucine residue at this position.

Table 1. Sequence Coverage for Tumor and Skin Genomes.*

Variable	Tumor	Skin
Sequencing runs — no.†	16.5	13.125
Haploid coverage	23.3×	21.3×
SNVs — no.	3,464,465	3,448,797
Concordance with dbSNP build 129 — no. (%)	3,053,215 (88.1)	2,992,069 (86.8)
High-quality heterozygous SNPs		
By array — no.	45,111	44,778
By sequence — no. (% of array)	44,442 (98.5)	43,629 (97.4)
High-quality homozygous rare SNPs		
By array — no.	28,295	27,735
By sequence — no. (% of array)	28,252 (99.8)	27,685 (99.8)

* The term dbSNP denotes a National Center for Biotechnology Information database of known DNA variants, SNP single-nucleotide polymorphism, and SNV single-nucleotide variant.
† A single sequencing run uses all eight lanes of an Illumina flow cell (see the Glossary).

TIER 1 MUTATIONS

The genes with tier 1 mutations and the consequences of these mutations are summarized in Table 2, and in Table 1 in the Supplementary Appendix. Both the *NPM*c insertion and the *NRAS* mutation have been described previously in AML genomes, and both are known to be relevant for pathogenesis.[3] Mutations in *IDH1* (encoding isocitrate dehydrogenase 1), which are predicted to affect the arginine residue at position 132, are found in malignant gliomas but have not been reported in patients with AML and are rare in other tumor types.[22-24] Variants of the nine other tier 1 genes are discussed in the Supplementary Appendix.

Each of the 10 point mutations was amplified from tumor and skin samples by means of PCR, and the DNA species carrying the variant allele was assayed by sequencing the PCR products with the use of the Illumina platform. The entire experiment was replicated with amplified genomic DNA, with excellent concordance for all samples (Fig. 3 in the Supplementary Appendix). The variant allele frequencies of the two insertions were determined by sequencing PCR products containing these mutations. The representation of all but two of the mutations — in chromosome 19 open reading frame 62 (*C19orf62*), an unannotated gene of unknown function, and *CEP170* — was approximately 50%, suggesting that all the mutations were heterozygous and present in nearly all the cells in the tumor sample (Fig. 2A). Ten of the 12 genes in tier 1 had probe sets on the Affymetrix U133 Plus 2.0 array, and 9 of 10 were detectably expressed (Table 1). We also assayed expression of the 10 nonsynonymous mutant alleles by means of reverse-transcriptase PCR, using amplicons designed to span introns, followed by sequencing and counting of the sequenced PCR products. Eight of the mutant alleles were detected

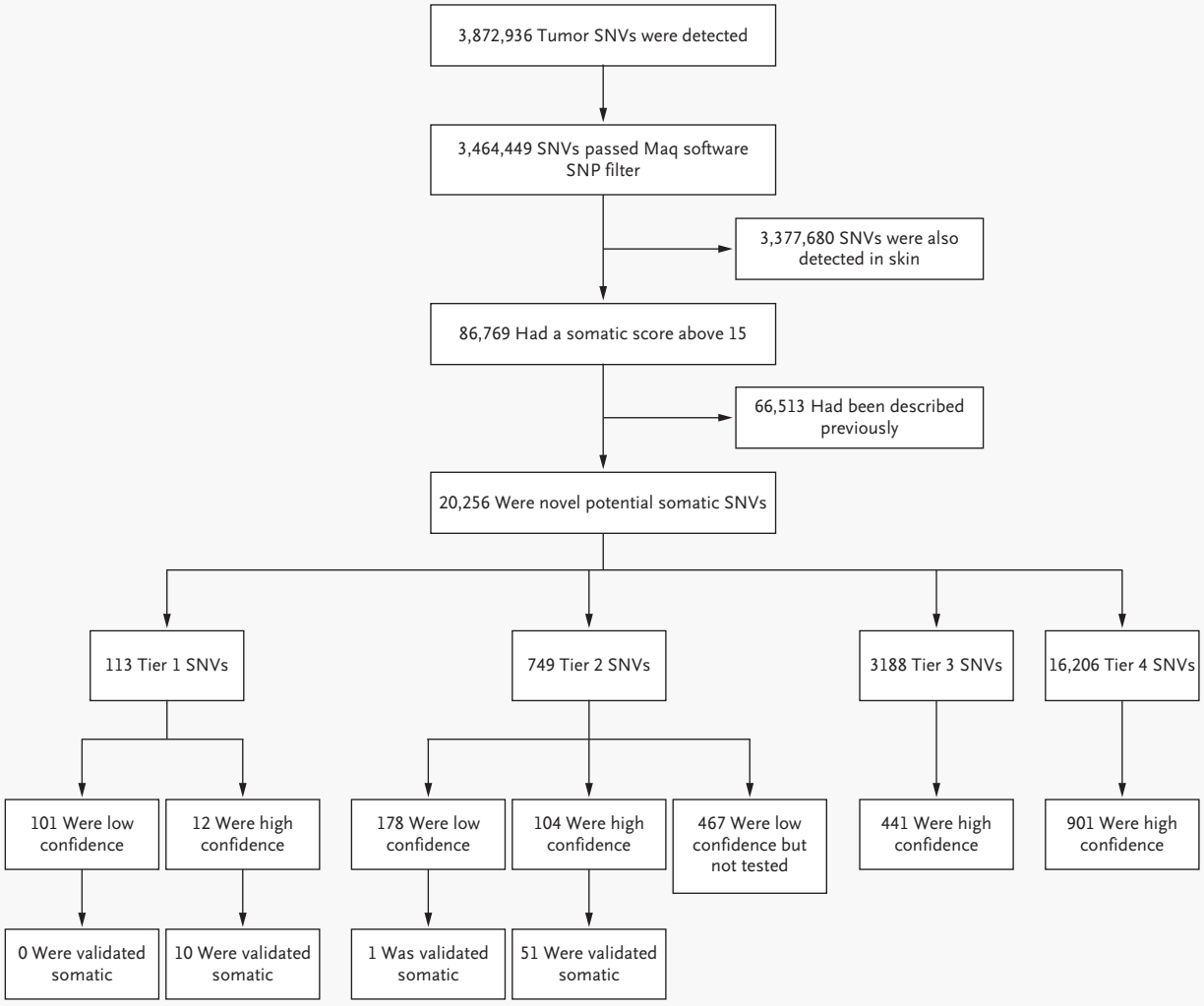

Figure 1. Flow Chart for Identification of Somatic Point Mutations in the Acute Myeloid Leukemia Genome.
Maq denotes Mapping and Assembly with Quality, SNP single-nucleotide polymorphism, and SNV single-nucleotide variant.

at frequencies of 35 to 85%. However, for two of the mutations (in *FREM2* and *IMPG2*) we did not detect complementary DNA carrying the variant allele (although we easily detected the wild-type allele), even though each variant was present in approximately 50% of the tumor DNA.

The individual bases that were mutated were highly conserved for 10 of the 12 variants, and all but 1 were found in highly conserved regions of the genome. The Sorting Intolerant from Tolerant (SIFT) algorithm (which gauges the likely effect of genic mutations on protein function) predicted that the mutations in *NRAS*, *IDH1*, *IMPG2*, and *ANKRD26* were deleterious.[25] The splice-site mutation at the 3′ end of intron 4 of *C19orf62* caused exon 5 to be skipped (data not shown).

We then genotyped the tier 1 mutations in 187 additional samples from patients with AML whose clinical characteristics have been described previously[26] (Table 2 in the Supplementary Appendix). The *NPMc* mutation was previously shown to be present in 43 of 180 samples (23.9%), and activating *NRAS* mutations were present in 17 of 182 samples (9.3%).[26] We observed mutations in *IDH1*, which were predicted to cause substitution of the arginine residue at position 132, in 16 of 188 samples: R132C in 8 samples, R132H in 7 samples, and R132S in 1 sample (Table 2 in the

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.