nature | Vol 456 | 6 November 2008 | doi:10.1038/nature07485

ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley[1,2,3,4]*, Elaine R. Mardis[2,3]*, Li Ding[2,3], Bob Fulton[3], Michael D. McLellan[3], Ken Chen[3], David Dooling[3], Brian H. Dunford-Shore[3], Sean McGrath[3], Matthew Hickenbotham[3], Lisa Cook[3], Rachel Abbott[3], David E. Larson[3], Dan C. Koboldt[3], Craig Pohl[3], Scott Smith[3], Amy Hawkins[3], Scott Abbott[3], Devin Locke[3], LaDeana W. Hillier[3,8], Tracie Miner[3], Lucinda Fulton[3], Vincent Magrini[2,3], Todd Wylie[3], Jarret Glasscock[3], Joshua Conyers[3], Nathan Sander[3], Xiaoqi Shi[3], John R. Osborne[3], Patrick Minx[3], David Gordon[8], Asif Chinwalla[3], Yu Zhao[1], Rhonda E. Ries[1], Jacqueline E. Payton[5], Peter Westervelt[1,4], Michael H. Tomasson[1,4], Mark Watson[3,4,5], Jack Baty[6], Jennifer Ivanovich[4,7], Sharon Heath[1,4], William D. Shannon[1,4], Rakesh Nagarajan[4,5], Matthew J. Walter[1,4], Daniel C. Link[1,4], Timothy A. Graubert[1,4], John F. DiPersio[1,4] & Richard K. Wilson[2,3,4]

Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adults in the United States each year. The treatment of this disease has changed little in the past two decades, because most of the genetic events that initiate the disease remain undiscovered. Whole-genome sequencing is now possible at a reasonable cost and timeframe to use this approach for the unbiased discovery of tumour-specific somatic mutations that alter the protein-coding genes. Here we present the results obtained from sequencing a typical acute myeloid leukaemia genome, and its matched normal counterpart obtained from the same patient's skin. We discovered ten genes with acquired mutations; two were previously described mutations that are thought to contribute to tumour progression, and eight were new mutations present in virtually all tumour cells at presentation and relapse, the function of which is not yet known. Our study establishes whole-genome sequencing as an unbiased method for discovering cancer-initiating mutations in previously unidentified genes that may respond to targeted therapies.

We used massively parallel sequencing technology to sequence the genomic DNA of tumour and normal skin cells obtained from a patient with a typical presentation of French–American–British (FAB) subtype M1 acute myeloid leukaemia (AML) with normal cytogenetics. For the tumour genome, 32.7-fold 'haploid' coverage (98 billion bases) was obtained, and 13.9-fold coverage (41.8 billion bases) was obtained for the normal skin sample. Of the 2,647,695 well-supported single nucleotide variants (SNVs) found in the tumour genome, 2,584,418 (97.6%) were also detected in the patient's skin genome, limiting the number of variants that required further study. For the purposes of this initial study, we restricted our downstream analysis to the coding sequences of annotated genes: we found only eight heterozygous, non-synonymous somatic SNVs in the entire genome. All were new, including mutations in protocadherin/cadherin family members (*CDH24* and *PCLKC* (also known as *PCDH24*)), G-protein-coupled receptors (*GPR123* and *EBI2* (also known as *GPR183*)), a protein phosphatase (*PTPRT*), a potential guanine nucleotide exchange factor (*KNDC1*), a peptide/drug transporter (*SLC15A1*) and a glutamate receptor gene (*GRINL1B*). We also detected previously described, recurrent somatic insertions in the *FLT3* and *NPM1* genes. On the basis of deep readcount data, we determined that all of these mutations (except *FLT3*) were present in nearly all tumour cells at presentation and again at relapse 11 months later, suggesting that the patient had a single dominant clone containing all of the mutations. These results demonstrate the power of whole-genome sequencing to discover new cancer-associated mutations.

AML refers to a group of clonal haematopoietic malignancies that predominantly affect middle-aged and elderly adults. An estimated 13,000 people will develop AML in the United States in 2008, and 8,800 will die from it[1]. Although the life expectancy from this disease has increased slowly over the past decade, the improvement is predominantly because of improvements in supportive care—not in the drugs or approaches used to treat patients.

For most patients with a 'sporadic' presentation of AML, it is not yet clear whether inherited susceptibility alleles have a role in the pathogenesis[2]. Furthermore, the nature of the initiating or progression mutations is for the most part unknown[3]. Recent attempts to identify additional progression mutations by extensively re-sequencing tyrosine kinase genes yielded very few previously unidentified mutations, and most were not recurrent[4,5]. Expression profiling studies have yielded signatures that correlate with specific cytogenetic subtypes of AML, but have not yet suggested new initiating mutations[6–8]. Recent studies using array-based comparative genomic hybridization and/or single nucleotide polymorphism (SNP) arrays, although identifying important gene mutations in acute lymphoblastic leukaemia[9,10] have revealed very few recurrent submicroscopic somatic copy number variants in AML (M.J.W., manuscript in preparation, and refs 11–13). Together, these studies suggest that we have not yet discovered most of the relevant mutations that contribute to the pathogenesis of AML. We therefore believe that unbiased whole-genome sequencing will be required to identify most of these mutations. Until recently, this approach has not been feasible because of the high cost of conventional

---

[1]Department of Medicine, [2]Department of Genetics, [3]The Genome Center at Washington University, [4]Siteman Cancer Center, [5]Department of Pathology and Immunology, [6]Division of Biostatistics, and [7]Department of Surgery, Washington University School of Medicine, St. Louis, Missouri 63108, USA. [8]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.

*These authors contributed equally to this work.

capillary-based approaches and the large numbers of primary tumour cells required to yield the necessary genomic DNA. 'Next-generation' sequencing approaches, however, have changed this landscape.

Our group has pioneered the use of whole-genome re-sequencing and variant discovery approaches using the Illumina/Solexa technology with the genome of the nematode worm *Caenorhabditis elegans* as a proof-of-principle[14]. This approach has distinct advantages in reduced cost, a markedly increased data production rate, and a low input requirement of DNA for library construction. In the present study, we used a similar approach to sequence the tumour genome of a single AML patient and the matched normal genome (derived from a skin biopsy) of the same patient. After alignment to the human reference genome, sequence variants were discovered in the tumour genome and compared to the patient's normal sequence, to the dbSNP database, and to variants recently reported for two other human genomes[15,16]; revealing new single nucleotide and small insertion/deletion (indel) variants genome-wide. Somatic mutations were detected in genes not previously implicated in AML pathogenesis, demonstrating the need for unbiased whole-genome approaches to discover all mutations associated with cancer pathogenesis.

Rationale for using the FAB M1 AML subtype for sequencing

Of the eight FAB subtypes of AML, M1 AML is one of the most common (~20% of all cases). No specific cytogenetic abnormalities or somatic initiating mutations have been identified for this subtype; in fact, about half of the patients with *de novo* M1 AML have normal cytogenetics[17–19]. The frequency of well-described progression mutations (for example, activating alleles of *FLT3*, *KIT* and *RAS*) is similar to that of other common FAB subtypes[5]. We therefore decided to sequence the genome of tumour cells derived from a patient with M1 AML, because so little is known about the molecular pathogenesis of this common subtype. The criteria used to select the sample are outlined in Supplementary Information.

Case presentation of UPN 933124

The case presentation is described in detail in the Supplementary Information. In brief, a previously healthy woman in her mid-50s presented suddenly with fatigue and easy bruisability, and was found to have a peripheral white blood cell count of 105,000 cells per microlitre, with 85% myeloblasts. A bone marrow examination revealed 100% myeloblasts with morphological features and cell surface markers consistent with FAB M1 AML (Supplementary Fig. 1). Cytogenetic analysis of tumour cells revealed a normal 46,XX karyotype. Although the patient experienced a complete remission with conventional therapies, she relapsed at 11 months and expired 24 months after her initial diagnosis was made. At relapse, the bone marrow had 78% myeloblasts, and contained a new clonal cytogenetic abnormality, t(10; 12) (p12; p13). Informed consent for whole-genome sequencing was subsequently obtained from her next of kin.

A typical M1 AML diploid genome and expression profile

The tumour sample from patient 933124 contained no somatic copy number changes at a resolution of ~5 kb (further confirmed on the NimbleGen 2.1M array platform, data not shown), and no evidence of copy number neutral loss-of-heterozygosity (LOH), indicating that the genome was essentially diploid at this level of resolution (see Supplementary Fig. 2). Further analysis of the 933124-derived tumour and skin samples showed 26 inherited copy number variants (that is, detected in both the tumour and skin samples). All but two of these had been previously reported in the Database of Genomic Variants (see Supplementary Table 1). All of the copy number variants detected in this genome were found in at least one other AML patient (89 other cases, mostly Caucasian, have been queried using the same SNP array platform), and all but one were found in at least one of the 160 Caucasian HapMap and Coriell samples that were studied on the same array platform (Supplementary Table 1).

To determine whether the tumour cells of 933124 were typical of M1 AML, we compared the expression signatures of 111 *de novo* AML cases using unsupervised clustering (Ward's method, see Supplementary Information). The expression profile of patient 933124 clustered with multiple other M1 (and M2) AML cases with normal cytogenetics, suggesting that the genetic events underlying the pathogenesis of this case are similar to those of other cases exhibiting normal cytogenetics (Supplementary Fig. 3).

Coverage depth of the tumour and skin genomes

Because most of the acquired mutations in cancer genomes have been shown to be heterozygous, the complete sequencing of a cancer genome requires the detection of both alleles at most positions in the genome[20]. We therefore designed sequence coverage metrics to define the point at which 90% diploid coverage had been reached. To minimize errors associated with any single platform or measurement, diploid coverage for this genome was assessed using a set of high-quality SNPs derived from two different SNP array platforms, Affymetrix 6.0 and Illumina Infinium 550K. For a SNP to be included in the high-quality set, the following criteria had to be satisfied: (1) identical genotypes were called from both assays at the same genomic positions, and (2) the resulting genotype was heterozygous. For the 933124 tumour genome, 46,494 heterozygous SNPs passed the above criteria and were defined as high-quality SNPs. For the skin samples, 46,572 high-quality SNPs were defined.

We performed 98 full runs on the Illumina Genome Analyser to achieve the targeted level of 90% diploid coverage as determined by coverage of the high-quality SNP set. Maq[21] was used to perform alignment, determine consensus, and identify SNVs within the 98 billion bases generated from the tumour genome (see Table 1). Maq predicted a total of 3.81 million SNVs (Maq SNP quality ≥ 15) in the tumour genome, including matching heterozygous genotypes for 91.2% of the 46,494 high-quality SNPs. When we lowered the Maq SNP quality cutoff to 0, 94.06% high-quality SNPs were predicted. Further investigation of Maq alignments revealed coverage for both alleles at a further 5.38% of the high-quality SNPs, but Maq did not predict a SNP or matching heterozygous genotype owing to insufficient depth or quality of coverage. Extra analysis revealed coverage at 46,484 of 46,494 high-quality SNPs for at least one allele (that is, 99.98% haploid coverage for the tumour genome).

We sequenced the genome of normal skin cells from the same patient to enable the identification of inherited sequence variants in the tumour genome. Our targeted diploid coverage goal for the skin-derived genome was 80%. We achieved this goal with only 34 Solexa runs (41.8 billion bases), using improved reagents and longer read lengths to attain 82.6% diploid and 84.2% haploid coverage (Table 1).

To begin evaluating the quantity and quality of the detected sequence variants in the tumour and skin genomes, we compared the overlap and uniqueness of this genome's variants with respect to the James D. Watson and J. Craig Venter genomes, and to dbSNP (v127; Fig. 1). Of the 3.68 million single nucleotide variants (SNVs; Maq SNP quality ≥ 15, excluding SNVs found on chromosome X) predicted by Maq in the tumour genome, 2.36 million were present in dbSNP, 2.36 million were detected in the skin genome (Fig. 1a), 1.50 million were detected in the Venter genome, and 1.58 million were found in the Watson genome (Fig. 1b). Ultimately, 1.70 million SNVs were unique to the 933124 tumour genome. On filtering the 933124 SNVs at different Maq quality values to determine the stability of results, we observed that the proportion of 933124 SNVs that also are in dbSNP increases from 63.9% to 69.48% when the Maq quality threshold score increases from 15 to 30, as expected.

Refining the detection of potential somatic mutations

Because the number of sequence variants initially detected by Maq was high, we developed improved filtering tools to effectively separate true variants from false positives. To this end, we generated an

Table 1 | Tumour and skin genome coverage from patient 933124

	Tumour	Skin
Libraries	4	3
Runs	98	34
Reads obtained	5,858,992,064	2,122,836,148
Reads passing quality filter	3,025,923,365	1,228,177,690
Bases passing quality filter	98,184,511,523	41,783,794,834
Reads aligned by Maq	2,729,957,053	1,080,576,680
Reads unaligned by Maq	295,966,312	138,276,594
SNVs detected with respect to hg18 (no Y)	3,811,115	2,918,446
SNVs (chr 1–22) detected with respect to hg18	3,681,968 (100.0%)	2,830,292 (100.0%)
SNVs also present in dbSNP	2,368,458 (64.3%)	2,161,695 (76.4%)
SNVs also present in Venter genome	1,499,010 (40.7%)	1,383,431 (48.9%)
SNVs also present in Watson genome	1,573,435 (42.7%)	1,456,822 (51.5%)
SNVs not in dbSNP/Venter/Watson	1,223,830 (33.2%)	591,131 (20.9%)
SNVs not in dbSNP/Venter/Watson/skin	925,200 (25.1%)	–
HQ SNPs	46,494 (100.0%)	46,572 (100.0%)
HQ SNPs where reference allele is detected	42,419 (91.2%)	38,454 (82.6%)
HQ SNPs where variant allele is detected	43,164 (92.9%)	39,220 (84.2%)
HQ SNPs where both alleles are detected	42,415 (91.2%)	38,454 (82.6%)

Assessments are shown of the haploid and diploid coverage of the tumour and skin genomes from AML patient 933124.
Chr, chromosome; hg18, human genome version 18; HQ, high quality.

experimental data set by re-sequencing Maq-predicted SNVs, randomly selecting a training subset and a test data set, whose annotations and features were submitted to Decision Tree C4.5 (ref. 22). This approach identified parameters that separated true variants from false positives, revealing that SNV-supporting read counts (unique on the basis of read start position and base position in supporting reads), base quality and Maq quality scores are chief determinants for identifying false positives. Implementing rules obtained from the Decision Tree analysis resulted in 91.9% sensitivity and 83.5% specificity for validated SNVs.

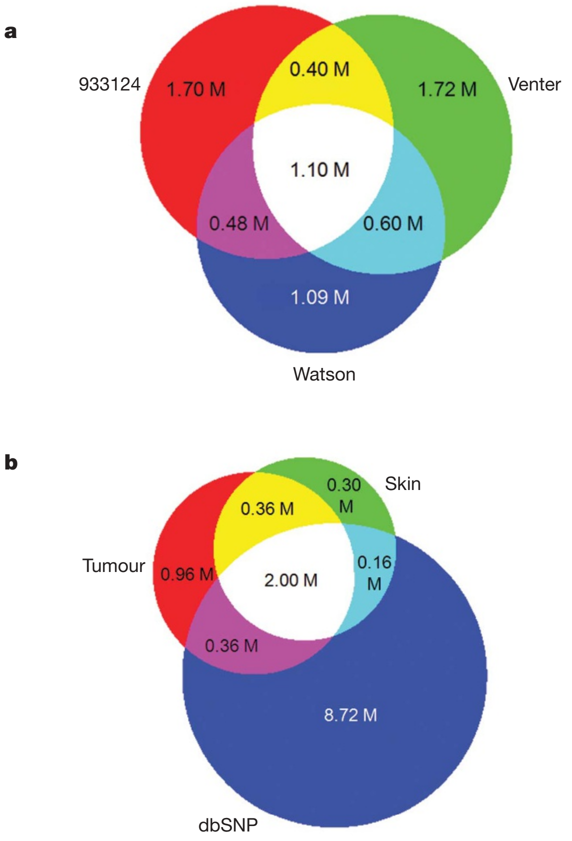

Figure 1 | Overlap of SNPs detected in 933124 and other genomes. a, Venn diagram of the overlap between SNPs detected in the 933124 tumour genome and the genomes of J. D. Watson and J. C. Venter. **b**, Venn Diagram of the overlap among the 933124 tumour genome, the skin genome and dbSNP (ver. 127). SNVs were defined with a Maq SNP quality ≥15.

Identification of somatic mutations in coding sequences

The patient had 3,813,205 sequence variants in her tumour genome, as defined by Maq scores of >15 (Table 1). Of these, 2,647,695 were supported by the Decision Tree analysis in the tumour genome, of which 2,584,418 (97.6%) were also detected in the skin genome (Fig. 2). The detailed algorithm for selecting putative somatic variants is described in Supplementary Information. Most of the 63,277 tumour-specific variants we detected were either present in dbSNP or were previously described in the Watson or Venter genomes (31,645), or occurred in non-genic regions (20,440). A total of 11,192 variants were located within the boundaries of annotated

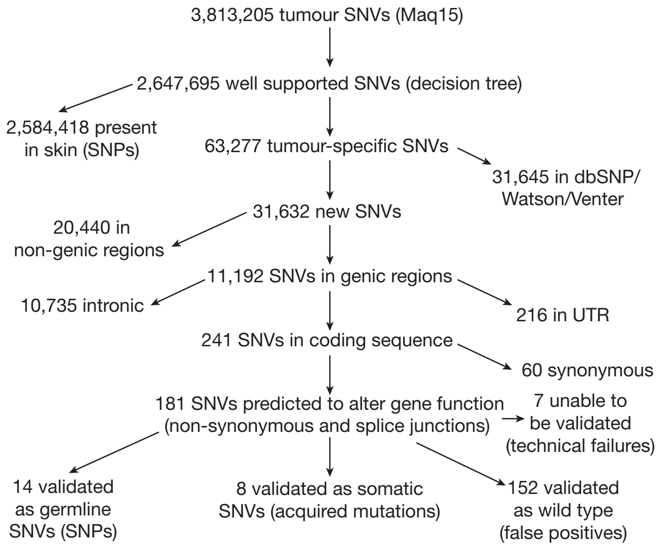
3,813,205 tumour SNVs (Maq15)
2,647,695 well supported SNVs (decision tree)
2,584,418 present in skin (SNPs)
63,277 tumour-specific SNVs
31,645 in dbSNP/ Watson/Venter
31,632 new SNVs
20,440 in non-genic regions
11,192 SNVs in genic regions
10,735 intronic
216 in UTR
241 SNVs in coding sequence
60 synonymous
181 SNVs predicted to alter gene function (non-synonymous and splice junctions)
7 unable to be validated (technical failures)
14 validated as germline SNVs (SNPs)
8 validated as somatic SNVs (acquired mutations)
152 validated as wild type (false positives)

Figure 2 | Filters used to identify somatic point mutations in the tumour genome. See text for details. UTR, untranslated regions.

genes; 216 of these variants were in untranslated regions, and 10,735 were in introns (but not involving splice junctions) and were not explored further in our analysis. Of the coding sequence variants, 60 were synonymous, and not further evaluated. The remaining 181 variants were either non-synonymous, or were predicted to alter splice site function. By sequencing polymerase chain reaction (PCR)-generated amplicons from the tumour and skin samples (and also from the relapse tumour sample obtained 11 months after the original presentation), we determined that 152 of these variants were false positive (that is, wild type) calls, 14 were inherited SNPs, and eight were somatic mutations in both the original tumour and the relapse sample (Table 2). Seven variants could not be validated, either because the regions involved were repetitive, or because all attempts to obtain PCR amplicons failed. All of the PCR-amplified exons from the eight genes containing validated somatic mutations were sequenced in 187 further cases of AML using samples from our discovery and validation sets[23]; no further somatic mutations were detected in these genes (data not shown). A description of how we estimated the false negative (12.45%) and false positive (0.06%) rates for SNVs over the entire genome is presented in Supplementary Information. Using these estimates, we can predict that very few somatic, non-synonymous variants were missed by our analysis of this deeply covered genome.

Defining mutation frequencies in the tumour sample

To better define the percentage of tumour cells that contained each of the discovered somatic mutations, we amplified each mutation-containing locus from non-amplified genomic DNA derived from the *de novo* and relapse tumour samples, and from the skin biopsy obtained at presentation. The resulting amplicons were sequenced using the Roche/454 FLX platform, and the frequency of reads containing the reference and variant alleles were defined (Fig. 3 and Table 3). Control amplicons containing a known heterozygous SNP in *BRCA2* (encoding N372H) and a homozygous SNP in *TP53* (encoding P72R) were analysed similarly. The *BRCA2* SNP yielded ~50% variant frequencies in the tumour and skin samples, whereas nearly 100% of the *TP53* alleles were variant in all three samples, as expected. Remarkably, all eight somatic SNVs were detected at ~50% frequencies in the primary tumour sample (100% blasts), and at ~40% frequencies in the relapse sample (78% blasts; if the variant frequencies are corrected for blast counts—that is, multiplied by 1.28—the frequencies at relapse also were ~50%). The NPMc (cytoplasmic nucleophosmin) mutation was also detected at a frequency of ~50%, but the *FLT3* internal tandem duplication (ITD) allele was only detected in 35.1% of the 454 reads at diagnosis and 31.3% at relapse, suggesting that the mutation was not present in all tumour cells at diagnosis or relapse.

Notably, the variant alleles also were detected at frequencies of ~5–13% in the skin sample. In retrospect, it is clear that the skin sample contained contaminating leukaemic cells, because the patient's white blood cell count at presentation was 105,000 per microlitre, with 85% blasts. This information was used to inform the Decision Tree analysis described above: we allowed high-quality tumour variants to move forward in the discovery pipeline if they were detected at a low frequency (two or fewer reads) in the skin sample, as defined by a binomial test.

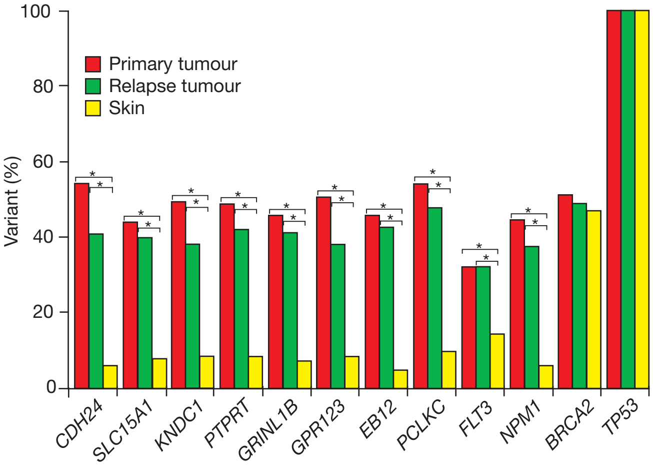

Figure 3 | Summary of Roche/454 FLX readcount data obtained for ten somatic mutations and two validated SNPs in the primary tumour, relapse tumour and skin specimens. The readcount data for the variant alleles in the primary tumour sample and relapse tumour sample are statistically different from that of the skin sample for all mutations ($P < 0.000001$ for all mutations, Fisher's exact test, denoted by a single asterisk in all cases). Note that the normal skin sample was contaminated with leukaemic cells containing the somatic mutations. The patient's white blood cell count was 105,000 (85% blasts) when the skin punch biopsy was obtained.

Detecting insertions and deletions (indels)

To discover small indels (<6 bp) from sequence reads (32–35 bp long), we started with a set of 236 million reads that were not confidently aligned by Maq to the reference genome. We applied Cross_Match and BLAT to identify gapped alignments that are unique in the genome. To detect indels longer than 6 bp, we developed a 'split reads' algorithm (see Supplementary Information) that aligns subsegments of reads independently to the genome, and computes a mapping quality for the derived gapped alignment on the basis of the number of hits and the quality of the bases. These efforts resulted in the identification of 726 putative small indels (1 to 30 bp in size) that occur in coding exons, 393 of which (54.2%) were found in dbSNP. After manual review, we selected a set of 28 putative somatic coding indels for validation using PCR-based dye terminator sequencing. Of these putative indels, 22 were validated but were found present in both tumour and skin (15 of these were in dbSNP), two were false positive calls, two had no coverage, and two were previously validated somatic insertions in *NPM1* (4 bp) and *FLT3* (30 bp).

Discussion

Here we describe the sequencing and analysis of a primary human cancer genome using next-generation sequencing technology. Our

Table 2 | Non-synonymous somatic mutations detected in the AML sample

Gene	Consequence	Type	Solexa tumour reads WT:variant	Solexa skin reads WT:variant	Conservation score of mutant base	Mutations in other AML cases*
CDH24	Y590X	Nonsense	9:9	16:0	0.998	0/187
SLC15A1	W77X	Nonsense	15:12	19:0	1.000	0/187
KNDC1	L799F	Missense	7:8	20:0	NA	0/187
PTPRT	P1235L	Missense	9:13	16:0	1.000	0/187
GRINL1B	R176H	Missense	15:10	14:0	NA	0/187
GPR123	T38I	Missense	11:11	13:0	NA	0/187
EBI2	A338V	Missense	7:12	18:2	1.000	0/187
PCLKC	P1004L	Missense	19:9	15:1	0.98	0/187
FLT3	ITD	Indel	18:12	8:0	NA	51/185
NPM1	CATG ins	Indel	36:6	33:0	NA	43/180

Ins, insertion; WT, wild type.
* Patient cohort defined in ref. 23.

Table 3 | 454 Readcount data for somatic mutations and known SNPs

		Primary AML (100% blasts)			Skin			Relapse (78% blasts)		
Gene	Consequence	Variant	Ref	Variant (%)	Variant	Ref	Variant (%)	Variant	Ref	Variant (%)
CDH24	Y590X	5672	4890	53.70	564	10358	5.16	3108	4599	40.33
SLC15A1	W77X	3817	4962	43.48	875	10773	7.51	4714	7173	39.66
KNDC1	L799F	4640	4848	48.90	770	8972	7.90	3883	6342	37.98
PTPRT	P1235L	998	1058	48.54	126	1489	7.80	350	493	41.52
GRINL1B	R176H	2211	2674	45.26	318	4461	6.65	1447	2070	41.14
GPR123	T38I	4618	4569	50.27	850	9751	8.02	3660	6057	37.67
EBI2	A338V	12750	15453	45.21	458	10088	4.34	2646	3627	42.18
PCLKC	P1004L	992	855	53.71	341	3153	9.76	705	773	47.70
FLT3	ITD	4220	7810	35.08	3475	23159	13.05	3870	8495	31.30
NPM1	CATG ins	1550	1974	43.98	143	2390	5.65	2303	3910	37.07
BRCA2	N372H	778	752	50.85	763	876	46.55	285	303	48.47
TP53	P72R	8989	1	99.99	8161	0	100.00	7914	6	99.92

The differences between variant frequencies in primary or relapse tumour samples and skin were highly significant for all somatic mutations ($P < 0.000001$, Fisher's exact test, one tailed). The *BRCA2* variant is a known heterozygous SNP in this genome, and the *TP53* variant is a known homozygous SNP.

patient's tumour genome was essentially diploid, and contained ten non-synonymous somatic mutations that may be relevant for her disease. These mutations affect genes participating in several well-described pathways that are known to contribute to cancer pathogenesis, but most of these genes would not have been candidates for directed re-sequencing on the basis of our current understanding of cancer. Hence, these results justify the use of next-generation whole-genome sequencing approaches to reveal somatic mutations in cancer genomes.

As we demonstrated in our re-sequencing of the genome of the *C. elegans* N2 Bristol strain[14], and again in this study, massively parallel short-read sequencing provides an effective method for examining single nucleotide and short indel variants by comparison of the aligned reads to a reference genome sequence. By sequencing our patient's tumour genome to a depth of >30-fold coverage, and gauging our ability to detect known heterozygous positions across the genome, we have produced a sufficient depth and breadth of sequence coverage to comprehensively discover somatic genome variants. A slightly lower coverage of the normal genome from this individual helped to identify nearly 98% of potential variants as being inherited, a critical filter that allowed us to more readily identify the true somatic mutations in this tumour. Our results strongly support the notion that hypothesis-driven (for example, candidate gene-based) examination of tumour genomes by PCR-directed or capture-based methods is inherently limited, and will miss key mutations. A further and important consideration is the demand for large amounts of genomic DNA by these techniques; this is a serious limitation when precious clinical samples are being studied. The Illumina/Solexa technology requires only ~1 μg of DNA per library, enabling the study of primary tumour DNA rather than requiring the use of tumour cell lines, which may contain genetic changes and adaptations required for immortalization and maintenance in tissue culture conditions.

A total of ten non-synonymous somatic mutations were identified in this patient's tumour genome. Two are well-known AML-associated mutations, including an internal tandem duplication of the *FLT3* receptor tyrosine kinase gene, which constitutively activates kinase signalling, and portends a poor prognosis[5,24,25], and a four-base insertion in exon 12 of the *NPM1* gene (NPMc)[26–28]. Both of these mutations are common (25–30%) in AML tumours, and are thought to contribute to progression of the disease rather than to cause it directly[29]. Notably, the frequency of the mutant *FLT3* allele in the primary and relapse tumour samples (35.08% and 31.30%, respectively) was significantly less than that of the other nine mutations ($P < 0.000001$ for both the primary and relapse samples). These data suggest that the *FLT3* ITD may not have been present in all tumour cells, and further, that it may have been the last mutation acquired.

The other eight somatic mutations that we detected are all single base changes, and none has previously been detected in an AML genome. Four of the genes affected, however, are in gene families that are strongly associated with cancer pathogenesis (including *PTPRT*, *CDH24*, *PCLKC* and *SLC15A1*). The other four somatic mutations occurred in genes not previously implicated in cancer pathogenesis, but whose potential functions in metabolic pathways suggest mechanisms by which they could act to promote cancer (including *KNDC1*, *GPR123*, *EBI2* and *GRINL1B*). We speculate about the roles of these mutations for the pathogenesis of this patient's disease in Supplementary Information.

The importance of the eight newly defined somatic mutations for AML pathogenesis is not yet known, and will require functional validation studies in tissue culture cells and mouse models to assess their relevance. Even though we could not detect recurrent mutations in the limited AML sample set that we surveyed, several lines of evidence suggest that these mutations may not be random, 'passenger' mutations. First, somatic mutations in this genome are extremely rare. The rarity of somatic variants, and the normal diploid structure of the tumour genome, argues strongly against genetic instability or DNA repair defects in this tumour. Conceptually, this result is further supported by the very small number of somatic mutations discovered in the expressed tyrosine kinases of AML samples[4,5]; genetic instability does not seem to be a general feature of AML genomes.

Second, on the basis of the equivalent frequencies of the variant and wild-type alleles for the mutations in the tumour genome (except for *FLT3* ITD), it is highly probable that all the mutations are heterozygous, and are present in virtually all of the tumour cells (Fig. 3). The latter suggests that these mutations may have all been selected for and retained because they are important for disease pathogenesis in this patient. Alternatively, all may have occurred simultaneously in the same leukaemia-initiating cell, but only a subset of the mutations (or an as-yet undetected mutation) is truly important for pathogenesis (that is, disease 'drivers' versus passengers). Although we suggest that the latter hypothesis is very unlikely on the basis of our current understanding of tumour progression, many more AML genomes will need to be sequenced to resolve this issue.

Third, the same mutations were detected in tumour cells in the relapse sample at approximately the same frequencies as in the primary sample. All of these mutations were therefore present in the resistant tumour cells that contributed to the patient's relapse, further suggesting that a single clone contains all ten mutations. Fourth, seven of the ten genes containing somatic mutations were detectably expressed in the tumour sample. *FLT3* and *NPM1* messenger RNAs were highly expressed in this tumour sample, as they are in virtually all AML samples. We detected mRNA from the *CDH24*, *SLC15A1* and *EBI2* genes on the Affymetrix expression array, whereas expression of *GRINL1B* and *PCLKC* were detected by PCR with reverse transcription (RT–PCR; data not shown). Expression of *KNDC1*, *PTPRT* and *GPR123* was not detected by either approach, but we cannot rule out expression of these genes in a small subset of tumour cells (for example, leukaemia-initiating cells). Furthermore, for the five point mutations where data are available, the mutated base is highly conserved across multiple species (Table 2).

Although we performed whole-genome sequencing on this cancer sample, we restricted our initial validation studies to the 1–2% of the genome that encodes genes. This raises the issue of whether sequencing the complementary DNA transcriptome of this tumour would have been a faster, cheaper and more efficient way of finding the mutations. Although this approach will undoubtedly be an important adjunct to whole-genome sequencing, there are several advantages to the approach we used: (1) coverage models for whole-genome libraries are at present better understood than for cDNA libraries, where transcript abundance can vary over many orders of magnitude; (2) even if the transcriptome had been sequenced, extensive characterization of the normal genome would have been required to distinguish inherited variants from somatic mutations; and (3) relevant non-synonymous mutations could be missed by cDNA sequencing, including mutations that result in RNA instability (splice variants, nonsense mutations), and/or mutations in genes expressed at low levels, or in only a small subset of tumour cells.

The additional non-coding and non-genic somatic variants in this genome (which we presently estimate at 500–1,000 on the basis of our calculated false positive and negative rates for non-synonymous mutations), will provide a rich source of potentially relevant sequence changes that will be better understood as more cancer genomes are sequenced.

In summary, we have successfully used a next-generation whole-genome sequencing approach to identify new candidate genes that may be relevant for AML pathogenesis. We cannot overemphasize the importance of parallel sequencing of the patient's normal genome to determine which variants were inherited; the identification of the true somatic mutations in this tumour genome would not have been feasible without this approach. Furthermore, until hundreds (or perhaps thousands) of normal genomes and other AML tumours are sequenced, the contextual relevance of the mutations found in this genome will be unknown. Nevertheless, the somatic mutations that we did find were neither predicted by the curation of previously defined cancer genes, nor by the study of this tumour using unbiased, high-resolution array-based genomic approaches. For AML and other types of cancer, whole-genome sequencing may therefore be the only effective means for discovering all of the mutations that are relevant for pathogenesis.

METHODS SUMMARY

Sequence end reads (average length for tumour genome, 32 bp, and for skin, 35 bp) were generated from Illumina/Solexa fragment libraries derived from the tumour or skin cells of patient 933124, using the Illumina Genome Analyser. The analysed reads were aligned to the human reference genome (NCBI Build 36) using Maq[21]. Coverage of the tumour and normal genomes was ascertained by comparison to the patient's heterozygous SNPs, established by compiling shared SNP calls monitored on the Affymetrix 6.0 and Illumina Infinium 550K genotyping platforms. We examined the Maq alignments by Decision Tree analysis to discover SNVs, as well as to identify copy number variants. Non-aligned reads were further analysed for indel discovery. For all putative variants, we attempted validation using custom PCR and capillary sequencing on the ABI 3730 platform. All validated somatic mutations were further analysed by Roche/454 sequencing of PCR-generated amplicons made from primary genomic DNA to compare readcounts of wild-type and mutant alleles in the primary tumour, skin and relapse tumour samples. A complete description of the AML case sequenced, and the materials and methods used to generate this data set are provided in the Supplementary Information.

Sequence variant deposition in dbGaP. High-quality sequence variants defined by Decision Tree (2,647,695 variants) will be deposited in the dbGaP database (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap) for review by approved investigators.

Received 28 May; accepted 16 September 2008.

1. Jemal, A. *et al.* Cancer statistics, 2008. *CA Cancer J. Clin.* **58,** 71–96 (2008).
2. Owen, C., Barnett, M. & Fitzgibbon, J. Familial myelodysplasia and acute myeloid leukaemia—a review. *Br. J. Haematol.* **140,** 123–132 (2008).
3. Mrozek, K., Marcucci, G., Paschka, P., Whitman, S. P. & Bloomfield, C. D. Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification? *Blood* **109,** 431–448 (2007).
4. Loriaux, M. M. *et al.* High-throughput sequence analysis of the tyrosine kinome in acute myeloid leukemia. *Blood* **111,** 4788–4796 (2008).
5. Tomasson, M. H. *et al.* Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood* **111,** 4797–4808 (2008).
6. Schoch, C. *et al.* Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. *Proc. Natl Acad. Sci. USA* **99,** 10008–10013 (2002).
7. Bullinger, L. *et al.* Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.* **350,** 1605–1616 (2004).
8. Valk, P. J. *et al.* Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med.* **350,** 1617–1628 (2004).
9. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
10. Mullighan, C. G. *et al.* *BCR–ABL1* lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453,** 110–114 (2008).
11. Raghavan, M. *et al.* Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res.* **65,** 375–378 (2005).
12. Paulsson, K. *et al.* High-resolution genome-wide array-based comparative genome hybridization reveals cryptic chromosome changes in AML and MDS cases with trisomy 8 as the sole cytogenetic aberration. *Leukemia* **20,** 840–846 (2006).
13. Rucker, F. G. *et al.* Disclosure of candidate genes in acute myeloid leukemia with complex karyotypes using microarray-based molecular characterization. *J. Clin. Oncol.* **24,** 3887–3894 (2006).
14. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans. Nat. Methods* **5,** 183–188 (2008).
15. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
16. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
17. Byrd, J. C. *et al.* Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). *Blood* **100,** 4325–4336 (2002).
18. Grimwade, D. *et al.* The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood* **92,** 2322–2333 (1998).
19. Mrozek, K., Heerema, N. A. & Bloomfield, C. D. Cytogenetics in acute leukemia. *Blood Rev.* **18,** 115–136 (2004).
20. Wendl, M. C. & Wilson, R. K. Aspects of coverage in medical DNA sequencing. *BMC Bioinformatics* **9,** 239 (2008).
21. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* doi:10.1101/gr.078212.108 (in the press).
22. Quinlan, J. R. *C4.5:Programs for Machine Learning* 302 (Morgan Kaufmann Publishers, 1993).
23. Link, D. C. *et al.* Distinct patterns of mutations occurring in de novo AML versus AML arising in the setting of severe congenital neutropenia. *Blood* **110,** 1648–1655 (2007).
24. Frohling, S. *et al.* Identification of driver and passenger mutations of FLT3 by high-throughput DNA sequence analysis and functional assessment of candidate alleles. *Cancer Cell* **12,** 501–513 (2007).
25. Levis, M. & Small, D. FLT3: ITDoes matter in leukemia. *Leukemia* **17,** 1738–1752 (2003).
26. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N. Engl. J. Med.* **352,** 254–266 (2005).
27. Thiede, C. *et al.* Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood* **107,** 4011–4020 (2006).
28. den Besten, W., Kuo, M. L., Williams, R. T. & Sherr, C. J. Myeloid leukemia-associated nucleophosmin mutants perturb p53-dependent and independent activities of the Arf tumor suppressor protein. *Cell Cycle* **4,** 1593–1598 (2005).
29. Kelly, L. M. *et al.* PML/RARα and FLT3-ITD induce an APL-like disease in a mouse model. *Proc. Natl Acad. Sci. USA* **99,** 8283–8288 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to our AML patients and their families, and to A. J. Siteman, whose generous and visionary gift provided the main funding source for this study. We thank G. Flance, D. Kipnis and K. Polonsky for their support, and C. Bloomfield, M. Caligiuri and J. Vardiman from the Cancer and Leukemia Group B for providing important AML samples for validation studies. We also thank the staff of The Genome Center at Washington University for their support of and their many contributions to this project, and H. Li of the Sanger Institute for assistance with the use of Maq. Further funding was provided by the National Cancer Institute

(T.J.L.), the National Human Genome Research Institute (R.K.W.), and the Barnes-Jewish Hospital Foundation (T.J.L.).

Author Contributions T.J.L. and R.K.W.: project conception and oversight. T.J.L. and E.R.M.: project leaders and analysis coordination. L.D.: supervised variant discovery and characterization, decision tree analysis. D.E.L.: decision tree analysis development. S.S.: automated variant detection by decision tree analysis. B.F.: variant validation oversight. B.F., P.M. and D.G.: Consed multiple sequence viewer development/programming. M.D.M.: auto-analysis and manual review of validation data. K.C.: copy number analysis, variant detection algorithm development. D.C.K.: indel detection algorithm development. K.C. and L.W.H.: indel detection. D.D.: IT and data management, data analysis automation leader. B.H.D.-S.: variant detection algorithm development. S.M. and M.T.: library optimization and construction. L.C.: data generation scheduling and oversight. R.A. and T.M.: variant validation assays. X.S.: variant annotation pipeline development. D.E.L.: variant annotation. J.R.O.: variant data management and pfam analysis. A.H.: validation assay design. C.P.: LIMS (Laboratory Information Management System) oversight. S.A.: LIMS trouble shooting/facilitation of variant detection. D.L.: data analysis. L.F.: production data oversight. T.W. and J.G.: data analysis algorithm development. V.M.: next-generation platform development. J.C. and N.S.: primary next-generation data production. A.C.: analysis oversight for mutation discovery. Y.Z.: manual review of sequence variants. R.E.R. and M.J.W.: comparative genomic hybridization analyses. R.E.R.: cDNA expression analyses. J.E.P.: gene expression array analysis. P.W., M.W., J.I. and S.H.: clinical data and specimen acquisition/processing/management. R.N.: bioinformatic analysis. J.B. and W.D.S.: statistical analysis. P.W., M.H.T., T.A.G., J.F.D. and D.C.L.: study design, execution and analysis. T.J.L., E.R.M., D.D., D.L., L.W.H., P.W., M.H.T., D.C.L., T.A.G., J.F.D. and R.K.W.: manuscript preparation.

Author Information The high-quality sequence variants have been deposited in the dbGaP database (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap) under the accession number phs000159.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.R.M. (emardis@wustl.edu).