Modern Pathology (2012) 25, 795–804

Targeted next generation sequencing of clinically significant gene mutations and translocations in leukemia

Eric J Duncavage[1], Haley J Abel[2], Philippe Szankasi[3], Todd W Kelley[4] and John D Pfeifer[1]

[1]*Department of Pathology and Immunology, Division of Anatomic and Molecular Pathology, Division of Laboratory and Genomic Medicine, Washington University, St Louis, MO, USA;* [2]*Division of Statistical Genetics, Washington University, St Louis, MO, USA;* [3]*ARUP Laboratories, Salt Lake City, UT, USA and* [4]*Department of Pathology, University of Utah, Salt Lake City, UT, USA*

Leukemias are currently subclassified based on the presence of recurrent cytogenetic abnormalities and gene mutations. These molecular findings are the basis for risk-adapted therapy; however, such data are generally obtained by disparate methods in the clinical laboratory, and often rely on low-resolution techniques such as fluorescent *in situ* hybridization. Using targeted next generation sequencing, we demonstrate that the full spectrum of prognostically significant gene mutations including translocations, single nucleotide variants (SNVs), and insertions/deletions (indels) can be identified simultaneously in multiplexed sequence data. As proof of concept, we performed hybrid capture using a panel of 20 genes implicated in leukemia prognosis (covering a total of 1 Mbp) from five leukemia cell lines including K562, NB4, OCI-AML3, kasumi-1, and MV4–11. Captured DNA was then sequenced in multiplex on an Illumina HiSeq. Using an analysis pipeline based on freely available software we correctly identified DNA-level translocations in three of the three cell lines where translocations were covered by our capture probes. Furthermore, we found all published gene mutations in commonly tested genes including *NPM1*, *FLT3*, and *KIT*. The same methodology was applied to DNA extracted from the bone marrow of a patient with acute myeloid leukemia, and identified a t(9;11) translocation with single base accuracy as well other gene mutations. These results indicate that targeted next generation sequencing can be successfully applied in the clinical laboratory to identify a full spectrum of DNA mutations ranging from SNVs and indels to translocations. Such methods have the potential to both greatly streamline and improve the accuracy of DNA-based diagnostics.
Modern Pathology (2012) **25,** 795–804; doi:10.1038/modpathol.2012.29; published online 16 March 2012

Keywords: acute myeloid leukemia; AML; clinical diagnostics; leukemia prognostics; next generation sequencing; targeted sequencing

The identification of recurrent cytogenetic findings and single gene defects such as those occurring in *FLT3*, *NPM1*, and *CEBPA* form the diagnostic basis for risk-adapted therapy in leukemia, and, in particular, acute myeloid leukemia (AML).[1–3] The 2008 World Health Organization classification of hematopoietic neoplasms recognizes seven such translocations in AML including t(8;21) (q22;q22), inv(16), t(15;17) (q22;q12), t(9;11) (p22;q23), t(6;9) (p23;q34), inv(3), and t(1;22) (p13;q13); and five translocations including t(9;22) (q34;q11.2), 11q23 rearrangements, t(12;21) (p13;q22), t(5;14) (q31.q32), and t(1;19) (q23;p13.3) in acute lymphoblastic leukemia.[4] Successful identification of these translocations and their variants, as well as an ever-expanding list of single gene defects is now required for an accurate classification of leukemia. Although the number of genes and translocations requiring evaluation grows with increasing knowledge of leukemia pathogenesis and treatment response, the methods by which such information is obtained have changed little in the last 10–20 years. Furthermore, as the number and complexity of tests required for each new leukemia diagnosis increases, so do the costs and associated burdens placed on the clinical laboratory.

Correspondence: Dr EJ Duncavage, MD, Assistant Professor, Department of Anatomic and Molecular Pathology, Washington University, 660 South Euclid Ave, St. Louis, MO 63110, USA.
E-mails: Eric.duncavage@wustl.edu and eduncavage@me.com
Received 18 May 2011; revised 27 October 2011; accepted 28 October 2011; published online 16 March 2012

In the last several years, new DNA sequencing methods collectively known as 'next generation sequencing' have greatly increased our knowledge of cancer genomes, yielding rapid and relatively low-cost data sets.[5–9] These studies have uncovered new prognostic markers in AML such as *DNMT3A* and *IDH1*, illustrating the importance of next generation sequencing in the research/discovery setting. In the clinical laboratory, however, such mutations are generally identified by more conventional methods such as PCR (and variants thereof), Sanger sequencing, and fluorescent *in-situ* hybridization (FISH). Although these methods are relatively simple to implement, the time required to optimize each makes it difficult to keep up with the rapid pace of research in the field. Although mutations in commonly tested genes involving AML prognosis such as *FLT3* and *NPM1* tend to have a limited spectrum of mutations, making them amenable to PCR-based testing, the broader range of mutations in genes such as *CEPBA*, *DNMT3A*, and *KIT* necessitates the use of direct sequencing, adding to the expense and turn-around time of testing.[5,10–12] The detection of recurrent, balanced chromosomal translocations is critical to leukemia prognosis and is generally done by FISH and G-banded cytogenetics, relying on direct visualization of DNA. Although FISH offers increased sensitivity over conventional cytogenetics, the identification of novel translocation partners, such as those involving the *MLL* locus, or variant breakpoints requires the use of multiple probes, again increasing the cost and complexity of testing.[13,14]

Although the acquisition of next generation sequencing data is now relatively straightforward (see Mardis[15] for an excellent review), its analysis can be extremely complicated and time consuming due not only to the volume of data (often >100 GB/run), but also the computational difficulty in aligning short reads (Figure 1). Next generation sequencing, as opposed to conventional Sanger sequencing, relies on massive parallelization of the sequencing process to generate large numbers of reads; however, these reads are generally much shorter (36–400 bp) than those obtained by Sanger sequencing. The increased efficiency of next generation sequencing

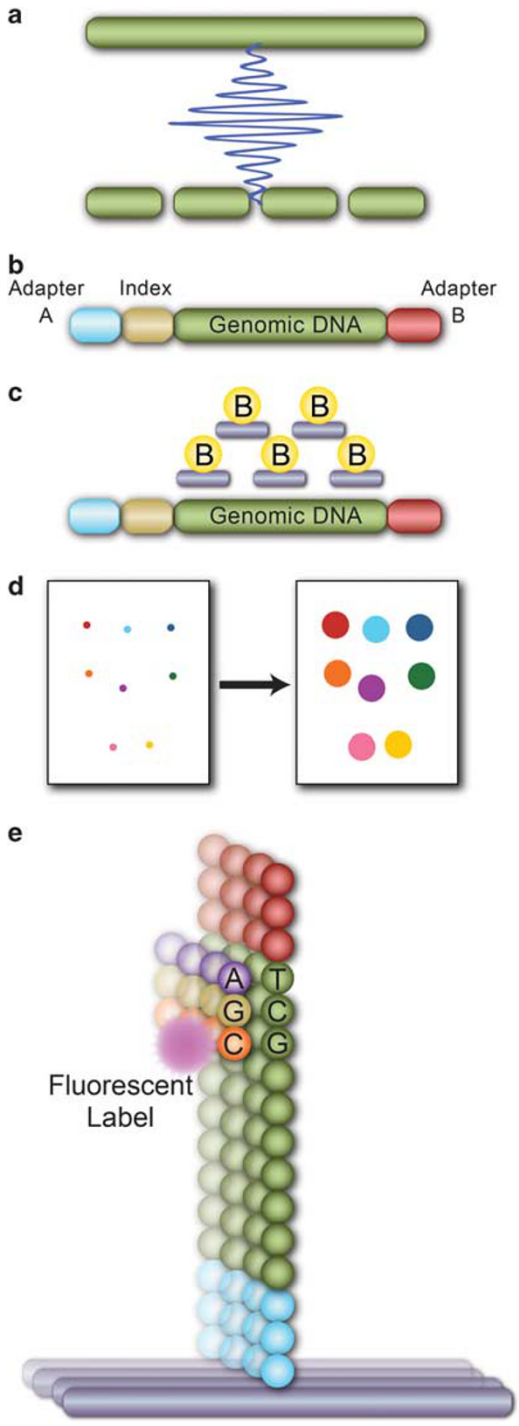

Figure 1 Hybrid capture enriched next generation sequencing. Although next generation sequencing encompasses a variety of technologies, each relies on massive parallelization of sequencing to achieve enormous throughput. In this example, Illumina sequencing is depicted. (**a**) Genomic DNA is sheared into small pieces (typically 300–500 bp) by sonication. (**b**) Sequencing adapters and sequencing indexes, the latter allowing for the identification of individual samples in pooled multiplexed data, are ligated to the sheared genomic DNA. (**c**) The prepared DNA (now called a library) is captured using biotinnylated cRNA oligomers specific for the region of interest. Following hybridization, this enriched DNA is eluted. (**d**) The enriched DNA libraries from multiple samples (each with unique index tags) are then loaded into a 'flow cell' containing immobilized oligomers with sequences complementary to the ligated library adapters. The library DNA then binds the surface and undergoes 'bridge amplification' to produce small colonies containing the amplified DNA library sequence. (**e**) DNA in the colonies is then sequenced using fluorescent, reversibly blocked nucleotides. Each nucleotide is labeled with a unique fluorophore, and, following incorporation of the complementary labeled base, each colony is scanned by a laser to determine the sequence of the last incorporated base. This process is performed in parallel over millions of colonies and repeated for each base, resulting in reads ranging from 36 to 150 bp depending on the chemistry and instrument used.

has greatly reduced the cost of sequencing to <0.00001 cents/base (compared with 1–5 cents/base by Sanger sequencing), allowing for the sequencing of whole human genomes for $5000.[16] Many software packages for the analysis for next generation sequencing data exist, including both freely available and commercial options. In this project, we relied solely on freely available software for both sequence alignment and downstream analysis. For the first step of sequence analysis we used both the Burroughs-Wheeler Aligner (BWA) and Novoalign, to perform alignments.[17,18] These software packages perform the computationally intensive task of aligning the short reads to the human reference genome. To identify sequence variations in the aligned data, we then used the Genome Analysis Tool Kit (GATK) Unified Genotyper software to identify single nucleotide variation (SNVs), and both the GATK and Pindel software to identify small and medium sized insertions and deletions (indels).[19,20] Finally, to find translocations we used the Breakdancer software package, which identifies discordant paired-end reads where one end maps to the targeted chromosome and the other to an alternate chromosome, and Slope, which identifies single end chimeric reads spanning the translocation boundary (Figure 2).[21,22] Although similar data can be obtained by commercial software packages,

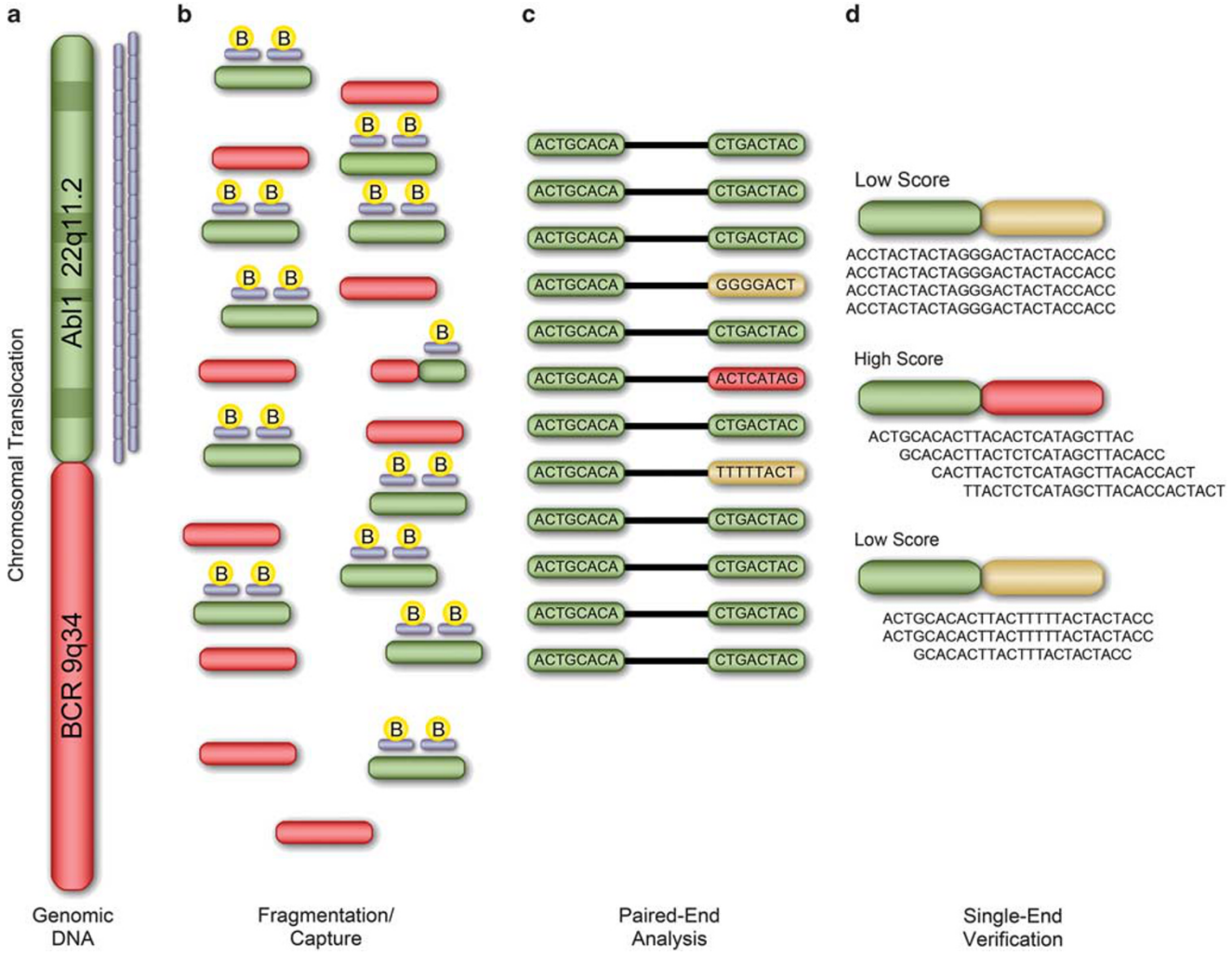

Figure 2 Overview of translocation identification by next generation sequencing. (**a**) Translocations occurring at the DNA level were identified by designing capture probes that 2× tiled across the both exons (dark green) and introns (light green) of gene partners commonly involved in translocations. In this example *ABL1* is captured, but its partner *BCR* is not. (**b**) Genomic DNA was then fragmented into ~300 bp pieces, library prepped, and captured. Genomic DNA containing sequences complementary to the *ABL1* (green)-specific biotin-labeled capture probes (blue), in this example, was enriched. Although most of the captured DNA represented contiguous areas of *ABL1*, regions with partial homology representing the actual DNA breakpoint (red and green) were also captured. (**c**) After aligning the sequence data, Breakdancer was used to identify paired reads in which one end of the paired-end read mapped to the targeted area (*ABL1*) and the other end did not (green/purple and green/red reads). (**d**) As the paired-end approach is subject to a high–false-positive rate due largely to DNA repeat regions, we employed a second level of filtering using Slope. Regions containing possible breakpoints were analyzed to find chimeric single end reads representing the actual translocation boundary (green/red). Chimeric reads mapping to repeat regions were removed based on their low score, and single high scoring hits identified.

publically available options offer a greater degree of customization and scalability.

Using targeted next generation sequencing, we propose a new paradigm in leukemia diagnostics in which prognostic information currently obtained by a variety of disparate methods can be acquired on a single platform with improved efficiency and markedly increased scalability. In this study, our goals were to prove that targeted next generation sequencing can be used to detect prognostically significant mutations and translocations, and that such data can be accurately analyzed using freely available software. As proof of concept, we leveraged the sequencing capacity of next generation sequencing to obtain high-fold coverage of genes with prognostic significance in leukemia via hybrid-capture enrichment (without gene-specific PCR amplification). In addition to correctly identifying SNVs, insertions, and deletions in commonly tested genes, we were able to reliably detect the t(15;17), t(9;22), and t(4;11) translocations in the NB4, K562, and MV4–11 cell lines, respectively, at the DNA level. Further, by capturing only one partner gene in a translocation (*RARA*, *ABL1*, *MLL*, etc) this methodology allows for the detection of novel partner genes or breakpoints, giving it a significant advantage over conventional break-apart FISH probes. Finally, we demonstrate the clinical utility of this methodology by identifying mutations and translocations in a patient-derived bone marrow sample.

Materials and methods

Cell Line and Patient Selection

To confirm the validity of this methodology we first sequenced DNA from five previously-characterized cell lines, including K562, NB4, MV4–11, kasumi-1, and OCI-AML3.[23–28] Bone marrow-derived DNA from a single anonymized patient with newly diagnosed AML and a known t(9;11) translocation was used to demonstrate the utility of the methodology on clinical material. The use of remnant patient samples for this study was approved by the University of Utah Investigational Review Board (IRB #7275).

Probe Design

120-bp cRNA Agilent SureSelect probes (Santa Clara, CA, USA) were designed to 2× tile across genes of interest in leukemia biology and prognosis (Table 1). To aid in the discovery of translocations, both introns and exons were covered by the probes with the exceptions of *RUNX1* and *MKL1,* in which only regions including exons/introns 1–3 and 1–5, respectively, were captured. Repeat masking was not performed to avoid the possibility of missing translocations occurring in repeat areas. The total size of the capture region was ~1.0 Mb.

Table 1 Gene capture coordinates (build 37 reference)

Gene	*Chromosome*	*Start*	*Stop*	*Size (bp)*
Translocations				
RUNX1	21	361 600 00	362 287 44	68 744
CBFB	16	67 063 050	67 134 956	71 906
RARA	17	38 474 497	38 513 894	39 397
MLL	11	117 812 415	117 901 146	88 731
NUP214	9	134 065 513	134 109 090	43 577
(MECOM)EVI1	3	168 801 287	168 851 758	50 471
MKL1	22	40 806 292	40816580	10 288
ABL	9	133 589 268	133 763 060	173 792
IL3	5	131 396 347	131 398 896	2549
E2A	19	1 609 293	1 650 286	40 993
TAL1	1	47 681 963	47 697 387	15 424
Mutations				
CEBPA	19	33 790 842	33 793 430	2588
NPMN	5	170 814 120	170 837 887	23 767
c-kit	4	55 524 095	55 606 879	82 784
FLT3	13	28 577 412	28 674 729	97 317
GATA1	X	48 644 982	48 652 715	7733
NOTCH1	9	139 396 889	139 408 003	11 114
KRAS	12	25 386 769	25 403 863	17 094
IDH1	2	209 100 954	209 119 806	18 852
LMO2	11	33 864 112	33 897 823	33 711

Capture and Sequencing

Approximately 2 μg of genomic DNA, extracted from cell lines or bone marrow, was fragmented to segments of between 250 and 500 bp using the Covaris S2 Sonolab (Covaris, Woburn, MA, USA). The resulting DNA was then end repaired and ligated to Illumina adapters (Illumina, San Diego, CA, USA) per the manufacturer's protocol. Sequence indexes were added to the cell line samples to permit all five cell lines to be sequenced in a single flow cell. Small fragments of <100 bp and unligated adapters were removed from the mix by AMPure purification (Agencourt Bioscience, Beverly, MA, USA). Sequencing libraries were then hybridized with SureSelect probes per the manufacturer's instructions. Streptavidin-coated paramagnetic beads were then added and allowed to bind the biotinylated capture probes. An external magnetic field was then applied and unbound DNA removed. The bound, captured DNA was finally eluted from the magnetic beads by digestion of the cRNA capture probes and purified. The enriched DNA was then amplified using universal primers targeting the paired-end adapters, clusters generated, and DNA sequenced on an Illumina HiSeq instrument with 2 × 101 bp reads in the case of cell line DNA, or an Illumina GAIIx with 2 × 60 bp reads for bone marrow-derived DNA.

Data Analysis

Base calls and quality scores were provided by the included Illumina software. The resulting FASTQ files were aligned to build 37 of the human reference genome (hg19) using Novoalign or BWA with default parameters.[17] Quality metrics including gene

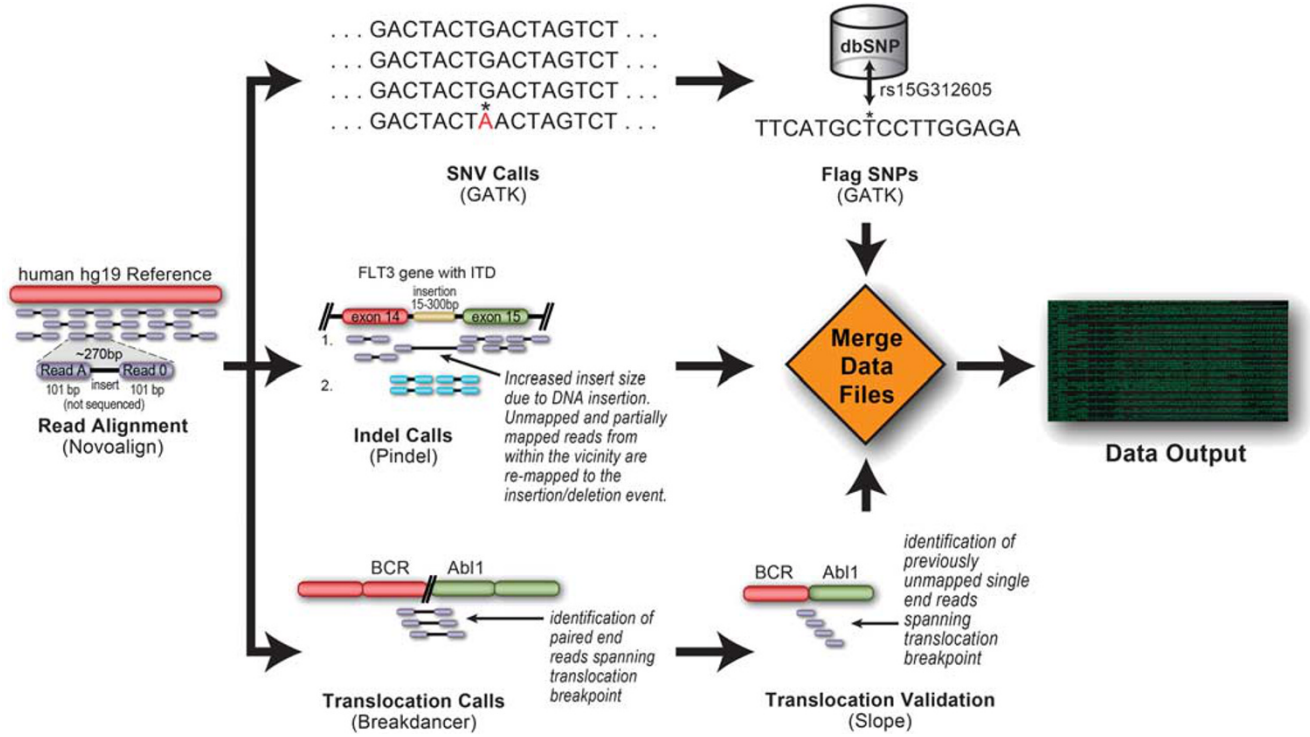

Figure 3 Data analysis pipeline. FASTQ files containing sequence and quality scores were output from the Illumina HiSeq and aligned to the human reference genome (build 37/hg19) using either BWA or Novoalign on a server cluster. The aligned data was then stored as a sorted BAM file and analyzed for SNVs, indels, and translocations. SNVs were called using the Unified Genotyper function of the GATK package. SNVs were further filtered by flagging known polymorphisms in dbSNP (build 130) and by removing SNVs occurring in non-coding regions. Small and medium size indels (<100 bp) were identified using Pindel and the GATK Indel Genotyper V2 software packages with default parameters. Indels occurring outside of coding regions or splice sites were ignored. Translocations were identified by first running Breakdancer to identify paired-end reads in which one end mapped to a gene in the capture region and the other did not. As this methodology is subject to considerable noise, largely because of sequence repeats and areas of homology, we then performed a second level of verification using Slope to find chimeric single end reads within the regions identified by Breakdancer. Finally, results from all three branches of the analysis pipeline were merged into single variant calling format (VCF) file.

coverage were calculated using BedTools.[29] SNVs were then called from the aligned sequence data using the UnifiedGenotyper in the GATK package.[30] The list of SNVs was then referenced against dbSNP (build 130) to flag known variants. The SNVs were further filtered by removing all SNVs occurring within non-coding regions. Indels events were identified using both Pindel and the GATK Indel Genotyper V2.0, whereas translocations were found by first using Breakdancer to identify clusters of paired-end reads in which the two members mapped to different chromosomes, and then verified using Slope to confirm the presence of chimeric single-end reads in the vicinity (10 kb) of the Breakdancer calls.[20–22] Default parameters were used with both programs. The above actions were combined using UNIX shell scripts to create an analysis pipeline (summarized in Figure 3).

Results

Quality Metrics

Total capture efficiency (percentage of total reads that mapped to the capture region) ranged from 5.7 to 14.9%. On average, each gene had 150 ± 69 × coverage (range 0–2215) in the five cell lines sequenced (Supplementary Figure 1 and Supplementary Table 1). Genes with the lowest fold coverage included *CEPBA* (41 ± 12 × average coverage) and *NOTCH1* (37 ± 12 × average coverage). These genes contain a higher percentage of regions with increased GC content (>70% GC) compared with the other captured genes, likely reducing the capture or library amplification efficiencies. By comparing GC content to fold coverage across the GC-rich *CEBPA* gene, we determined that optimal coverage is achieved with a GC content <70% (averaged over 200 bp increments) (Supplementary Figure 2). GC content >70% results in a considerable decrease in fold coverage, quickly leading to areas of zero coverage.

Identification of SNVs and Indels

We employed a data analysis pipeline comprised of freely available software to align and analyze next generation sequencing data. Single base substitutions were called using the GATK Unified Genotyper

Table 2 Next generation sequencing findings in commonly tested genes

	Reported translocation	*Captured gene location*	*Partner gene location*	FLT3 *ITD*	FLT3 *D835*	NPM1	KIT
OCI-AML3	N/A			–	–	+	–
NB-4	t(15;17) PML-RARA	Chr17: 38502182	Chr15: 74326368	–	–	–	–
Kasumi-1	t(8:21) AML-ETO	Undetermined[a]	Undetermined[a]	–	–	–	+
K562	t(9;22) BCR-ABL	Chr9: 133607147	Chr22: 23632742	–	–	–	–
MV4-11	t(4;11) MLL-AF4	Chr11: 118353675	Chr4: 88009085	+	–	–	–
Case 100	t(9;11) MLL-AF9	Chr11: 118354598	Chr9: 20355515	–	–	–	–

[a]The translocation breakpoint in this case occurred outside of the RUNX1 capture region (truncated because of design limitations) and was not sequenced.

and then compared with dbSNP to exclude known variants. Results were further filtered to remove SNVs occurring in non-coding regions. To identify indels, we used both Pindel and the GATK Indel Genotyper V2, and found that although the two methods produced identical results in the majority of calls, there were differences. For example, we compared both methods using sequence data from the 97-kb (including introns) *FLT3* gene in MV4–11 AML cell line. Using default parameters the GATK Indel Genotyper V2 identified 41 indels in the region whereas Pindel identified 66. Of these calls, the majority (35) were shared between the two programs; however, the 30-bp *FLT3* internal tandem duplication (ITD) (the only orthogonally validated indel) was identified by Pindel only.

By applying this methodology, we correctly identified all published indels and single base mutations (SNVs) occurring within the five cell lines (Table 2). For example, we identified previously published mutations, including the 4-bp type A *NPM1* exon 12 insertion in OCI-AML3, the 30-bp exon 14 *FLT3* ITD in MV4–11, and the D822K *KIT* mutation in exon 17 of kasumi-1. Coverage across gene regions of clinical interest such as the *NPM1* mutation, *FLT3* ITD, and *KIT* mutations were then determined to establish a coverage estimate for the identification of such mutations and averaged across all five cell lines (Figure 4a and b). Average coverage within a 1-kb region flanking the *NPM1* insertion and *FLT3* ITD were $174\times$ and $106\times$, respectively. Coverage within the *CEBPA* gene was too low to permit reliable identification of mutations. Other common mutations such as rs16754 in *WT1* and rs11554137 in *IDH1* were not observed.[31] Additional mutations of unknown clinical significance are included in Supplementary Table 2.

Identification of Translocations

Recurrent translocations were identified in three of the three cell lines in which one translocation partner was within the targeted capture area. Of the remaining two cell lines sequenced, the t(1;18)(p11;q11) described in OCI-AML3 is not considered a recurrent translocation and was not covered by the capture probes; the *RUNX1* breakpoint in kasumi-1 occurred outside of the capture region (the *RUNX1* capture region was truncated because of design limitations) and consequently was not sequenced.[23,32] Both the major and reciprocal translocations were identified in the K562 t(9;22) and NB4 t(15;17) cell lines, whereas only the major translocation was identified in the MV4–11 t(4;11) cell line. All translocations identified from next generation sequencing data were then validated by PCR and Sanger sequencing (verified breakpoint contigs are included in Supplementary Table 3).[33] Furthermore, primers for the specific translocations were tested on additional AML cases and showed no evidence of non-specific amplification, ruling out the possibility of incorrectly identifying repeat regions as translocations. Coverage in the translocation areas was then calculated to determine the necessary coverage required for finding translocations. In the case of the five successful translocation identifications (in three cell lines) average coverage was $176\times$ within a 1-kb region of the breakpoint (Figure 4c and d).

By applying a second level of software analysis beyond paired-end read mapping, only a single candidate translocation was identified in each cell line and all candidates proved to be correct. We initially applied Breakdancer to aligned data to find candidate translocation events based on paired-end reads in which one end mapped to a targeted translocation region (*ABL1*, *EVI1*, *CBFB*, *MKL1*, *MLL*, *NUP214*, *RARA*, and *RUNX1*) and the mate did not. We then evaluated areas surrounding the coordinates of putative translocations with the Slope software to verify the presence of chimeric single end reads spanning the translocation breakpoint (Figure 5). For each cell line, Breakdancer identified an average of 75 high-scoring (score >90) putative translocations, however, with the application of single-end read verification this number was reduced to one translocation in MV4–11 and two translocations (major and reciprocal) in K562 and NB4.

Patient Case

To demonstrate that the same methodology can be applied to lower quality clinical diagnostic material,

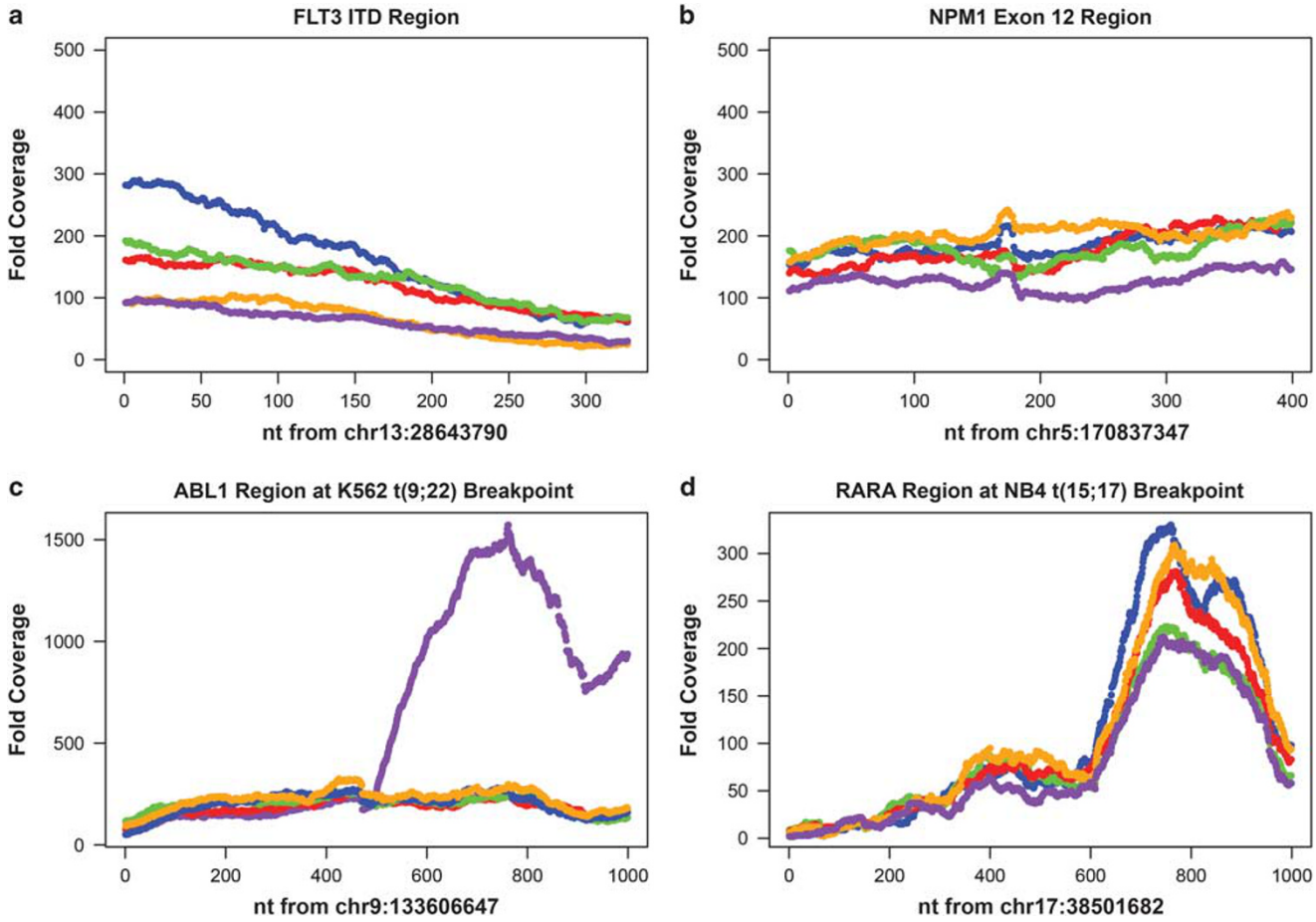

Figure 4 Coverage of clinically important genes and translocation areas. Blue: NB4, red: MV4–11, green: OCI-AML3, purple: K562, and orange: kasumi-1. (**a**) *FLT3* coverage for all cases in the region of the *FLT3* ITD. Average coverage was 116-fold. (**b**) *NPM1* coverage in the area surrounding exon 11 that generally harbors insertions. Average coverage was 174-fold. (**c**) Coverage of 1 kb in *ABL1* gene flanking the translocation site (middle) in K562 cells. Note that coverage in K562 increases dramatically at the breakpoint, likely indicating a copy number change in this region. (**d**) In contrast, the 1-kb area surrounding the t(15;17) RARA breakpoint in NB4 cells show no evidence of copy number change.

we performed the same capture/sequencing steps using DNA derived from the non-enriched bone marrow of an anonymized patient with newly diagnosed AML. Previous testing showed an *MLL* rearrangement using breakapart FISH probes, and the partner was subsequently confirmed to be *AF9,* t(9;11), using a panel of common *MLL* fusion partner probes. Testing for common AML gene mutations including *NPM1*, *FLT3* (ITDs, and D835), *CEBPA*, and *KIT* (exons 8 and 17) was performed at ARUP Laboratories and all produced negative results (data not shown). Next generation sequencing of the captured DNA demonstrated no clinically significant mutations (no previously described pathogenic mutations in coding regions), consistent with the findings obtained by conventional methods. We then used the same data analysis pipeline described for the cell line data to identify the t(9;11) *MLL-AF9* translocation. A single translocation event occurring within intron 8 (chr11: 118354598) of MLL and intron 23 (chr9: 20355515) of *AF9*/*MLLT3* was identified. This breakpoint was verified by standard PCR and Sanger sequencing, and was not present by PCR in other AML cases. Coverage analysis showed similar findings to the cell line data; genes including *KIT*, *IL3*, and *RUNX1* showed the highest fold coverage whereas genes such as *CEBPA* and *NOTCH1* exhibited low coverage. Coverage within the *MLL* gene at the translocation area was $\sim 300\times$.

Discussion

Here we demonstrate a relatively simple method based on hybrid capture and next generation sequencing for the simultaneous identification of single gene mutations (including SNVs and indels) and translocations in leukemia. Using DNA from cell lines with previously characterized findings, we identified all published mutations occurring within genes on the capture panel, without false positives. We further identified translocations in three of three cell lines and in one patient sample (one of one) by analyzing both paired-end and single-end read data.

a Discordant Paired End Reads (Breakdancer)

Chr:A	Position A	Chr:B	Position B	Score
chr11	51579705	ChrX	58561383	99
chr11	51579705	ChrX	61720890	99
chr16	33974645	chr21	10702156	99
chr2	92305681	chr18	18520129	99
chr9	78934308	chr19	53080527	99
chr9	134074576	chr22	17300016	99
chr9	133607341	chr22	23632504	99

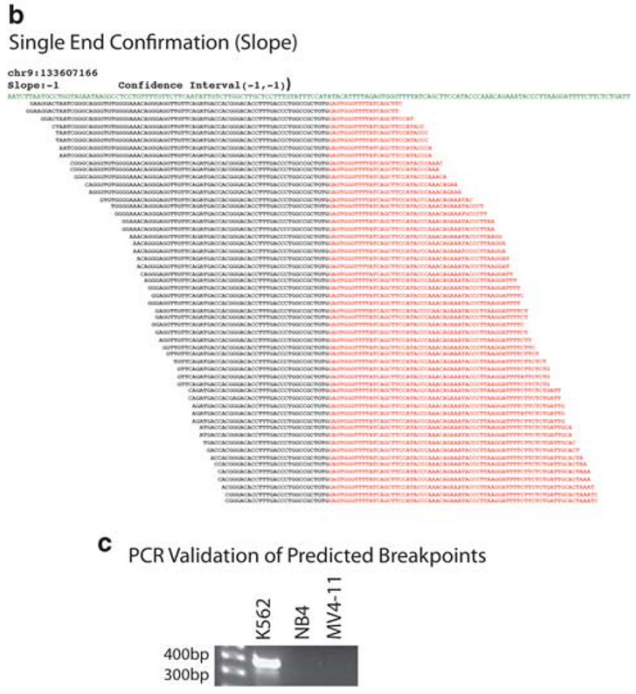

Figure 5 Identification of a t(9;22) translocation in the K562 cell line. (**a**) As only the *ABL1* gene was directly targeted, we first identified reads in which one paired read mapped to *ABL1* and the other did not using Breakdancer. This produced a table of ~150 hits (partial list shown), the majority representing matches to repeat regions. Of the 150 hits, 11 (including the actual breakpoint) had maximal scores of 99. (**b**) Breakdancer coordinates were then passed to SLOPE to perform single end chimeric read verification. Only a single high-scoring hit was identified using this methodology. (**c**) These findings were subsequently verified by constructing PCR primers flanking the proposed translocation and then Sanger sequencing the products. To demonstrate specificity we attempted to amplify the same translocation from other cell lines as shown.

Coverage data indicated that 177× coverage was sufficient for reliable translocation detection, whereas slightly lower-fold coverage, ~150×, allowed for detection of SNVs and indels. Together these data demonstrate that targeted next generation sequencing is a viable clinical laboratory method that has the potential to replace a number of conventional methods such as Sanger sequencing, capillary-based sizing, and FISH for the detection of clinically significant DNA mutations.

Targeted clinical next generation sequencing methods offer considerable advantages compared with standard laboratory methods; however, some obstacles remain before clinical next generation sequencing enters mainstream use. We and others have found that next generation sequencing data analysis is the greatest impediment to use.[34] For example, the single HiSeq lane containing the five indexed cell lines produced 28 GB of raw data, took ~50 CPU hours to align, and required some basic knowledge of bioinformatics to analyze. We relied on freely available, peer-reviewed tools such as GATK, Pindel, Breakdancer, and Slope to identify mutations within aligned data and used perl scripts to filter out mutations of clinical interest. In our hands, these methods showed a high sensitivity and specificity in detecting clinically significant SNVs and indels in cell lines where the mutations were present in at least 50% of the sequenced cells. We note that the detection of medium sized indels, especially the *FLT3* ITD, is difficult. Although the *FLT3* ITD was identified in MV4–11 cells using Pindel it could not be detected by the more general tools included in the GATK. This fact further highlights that multiple software tools will likely be required to analyze the full spectrum of mutation in clinical next generation sequencing samples. The limit of sensitivity remains to be determined for cases in which pathogenic SNVs and indels in leukemic cells are diluted by a larger population of normal cells, such as in the case of some AML with low-blast counts, or post-chemotherapy samples. In theory, blasts may be enriched by initial flow cytometry sorting before capture, but such methods remain untested.

We used simple 2× tiling without repeat masking in designing capture probes that resulted in an average of 9.8% of reads mapping to the targeted reference. This level of enrichment proved adequate for the detection of SNVs and indels in most genes; however, several genes such as CEBPA exhibited poor capture. We found that regions with high-GC content (>70%) exhibited low coverage. Optimal coverage was seen in the range of 20–69% GC content. Clearly the efficient capture of GC rich areas is problematic with current methods. Possible solutions including re-designing probe spacing or sizing to better match melting temperature content among probes and changing PCR conditions used in library construction.

Using this methodology we identified DNA-level translocations by capturing only one partner gene, thereby allowing for the detection of variant translocations that may be difficult to identify by FISH. For example, we captured only the *MLL* gene, but were able to identify both the t(4;11) translocation in MV4–11 cells as well as a patient t(9;11) translocation with single base accuracy by looking for sequencing reads that spanned non-adjacent segments of the genome. This method essentially exploits the 'off-target' reads, or areas of 'shoulder coverage', inherent to capture-based methods, but not present in PCR-based enrichment strategies. Although this approach worked successfully in the majority of cases, it should be noted that translocation sites often contain areas of homology or repeats

that would make paired end or chimeric reads difficult to align.[35] Therefore, even with sufficient coverage this method may not be capable of identifying all translocation events, and additional methodologies such as the inclusion of mate pair libraries may be required.[36]

Unlike current methods employed in the clinical molecular oncology laboratory, target-capture-based next generation sequencing offers greatly increased scalability, requires less technician labor, and is becoming less expensive. For example, mutations in many genes recently implicated in leukemia prognosis, such as *CEBPA* and *DMNT3A* occur throughout the coding region, and require sequencing of multiple PCR products and/or exons for full evaluation.[5,37] Although cumbersome by Sanger sequencing, obtaining sequence data over large coding areas and across multiple exons is fairly simple using target capture, requiring only the *in silico* design of sequence-specific capture probes and minimal optimization for most genes (depending on GC content as described earlier). It is also straightforward to add additional genes to a capture panel to accommodate new prognostic and diagnostic markers. Finally, the procedures involved in capture-based next generation sequencing can be highly automated and require minimal technician time for set up and running compared with standard methods.

In summary, we present proof of concept data showing that targeted next generation sequencing can be used in the clinical setting to detect prognostically significant mutations and translocations in leukemia. This methodology has the potential to replace a variety of more labor-intensive methods currently used to detect gene mutations and translocations in the clinical laboratory. In addition, we present a framework for the automated analysis of clinical next generation sequencing data using freely available software tools. Although we present data demonstrating the clinical utility of targeted next generation sequencing in leukemia, similar methods could be applied to solid tumors, for the simultaneous detection of both gene mutations and translocations, such as *ERG-TMPRSS2* fusions in prostate cancer or *EML4-ALK* fusions in non-small cell lung cancer.[38,39] By identifying such rearrangements at diagnosis with single base accuracy from DNA, patient-specific primers can be created to monitor for subsequent disease recurrence by more sensitive real-time PCR methods.[40] This method could in theory allow for disease monitoring in solid tumors similar to the use of *BCR-ABL1* quantitative PCR in chronic myelogenous leukemia. Further, we have previously used a similar approach to identify viral insertion sites with DNA derived from formalin-fixed paraffin-embedded tissue, suggesting that the methodology described herein may be amenable to archival material.[41] Finally, these methods are largely independent of sequencing platform and should be adaptable to other instrumentation depending on the particular needs of a laboratory.

Acknowledgements

Funding for this project was provided by internal funds from the Washington University of Department of Pathology. We thank Dr Sarah South at ARUP Laboratories for providing anonymized bone marrow DNA, Mark Johnson at the WU GTAC for assistance using the high-performance computer cluster, and Jon Armstrong of Cofactor Genomics in St Louis, MO for providing sequencing services. We also thank Jonathan Klein, MD of Washington University for his critique of this manuscript and David Spencer, MD, PhD for the use of his GATK analysis scripts.

Disclosure/conflict of interest

The authors declare no conflict of interest.

References

1 Walter MJ, Graubert TA, Dipersio JF, *et al.* Next-generation sequencing of cancer genomes: back to the future. Per Med 2009;6:653.
2 Smith ML, Hills RK, Grimwade D. Independent prognostic variables in acute myeloid leukaemia. Blood Rev 2010;25:39–51.
3 Betz BL, Hess JL. Acute myeloid leukemia diagnosis in the 21st century. Arch Pathol Lab Med 2010;134:1427–1433.
4 Arbor DA, Brunning RD, Le Beau MM, *et al.* Acute myeloid leukemia with recurrent genetic abnormalities. In: Swerdlow SH, Campo E, Harris NL, *et al.* (eds). WHO Classification of Tumors of Haematopoietic and Lymphoid Tissue Vol., 4 edn. IARC Press: Geneva, 2008 pp. 110–123.
5 Ley TJ, Ding L, Walter MJ, *et al.* DNMT3A mutations in acute myeloid leukemia. N Engl J Med 2010;363: 2424–2433.
6 Lee W, Jiang Z, Liu J, *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 2010;465:473–477.
7 Ley TJ, Mardis ER, Ding L, *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008;456:66–72.
8 Mardis ER, Ding L, Dooling DJ, *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 2009;361:1058–1066.
9 Stephens PJ, McBride DJ, Lin ML, *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 2009;462:1005–1010.
10 Taskesen E, Bullinger L, Corbacioglu A, *et al.* Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. Blood 2010;117:2469–2475.
11 Patel KP, Ravandi F, Ma D, *et al.* Acute myeloid leukemia with IDH1 or IDH2 mutation: frequency and clinicopathologic features. Am J Clin Pathol 2010;135:35–45.
12 Paschka P, Marcucci G, Ruppert AS, *et al.* Adverse prognostic significance of KIT mutations in adult acute myeloid leukemia with inv(16) and t(8;21): a Cancer and Leukemia Group B Study. J Clin Oncol 2006;24: 3904–3911.

13 Meyer C, Kowarz E, Hofmann J, *et al.* New insights to the MLL recombinome of acute leukemias. Leukemia 2009;23:1490–1499.
14 Welch JS, Westervelt P, Ding L, *et al.* Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 2011;305:1577–1584.
15 Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008;9:387–402.
16 Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Large Scale Genome Sequencing Program. 2011, Available from http://www.genome.gov/sequencingcosts/ (accessed 20 September 2011).
17 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–1760.
18 Hercus C. Novoalign. 2009 Available from http://www.novocraft.com(accessed 2 May 2011).
19 Depristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–498.
20 Ye K, Schulz MH, Long Q, *et al.* Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865–2871.
21 Chen K, Wallis JW, McLellan MD, *et al.* Break Dancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009;6:677–681.
22 Abel HJ, Duncavage EJ, Becker N, *et al.* SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. Bioinformatics 2010;26:2684–2688.
23 Quentmeier H, Martelli MP, Dirks WG, *et al.* Cell line OCI/AML3 bears exon-12 NPM gene mutation-A and cytoplasmic expression of nucleophosmin. Leukemia 2005;19:1760–1767.
24 Quentmeier H, Reinhardt J, Zaborski M, *et al.* FLT3 mutations in acute myeloid leukemia cell lines. Leukemia 2003;17:120–124.
25 Gale RP, Canaani E. An 8-kilobase abl RNA transcript in chronic myelogenous leukemia. Proc Natl Acad Sci USA 1984;81:5648–5652.
26 Lanotte M, Martin-Thouvenin V, Najman S, *et al.* NB4, a maturation inducible cell line with t(15;17) marker isolated from a human acute promyelocytic leukemia (M3). Blood 1991;77:1080–1086.
27 Megonigal MD, Rappaport EF, Wilson RB, *et al.* Panhandle PCR for cDNA: a rapid method for isolation of MLL fusion transcripts involving unknown partner genes. Proc Natl Acad Sci USA 2000;97:9597–9602.
28 Larizza L, Magnani I, Beghini A. The Kasumi-1 cell line: a t(8;21)-kit mutant model for acute myeloid leukemia. Leuk Lymphoma 2005;46:247–255.
29 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–842.
30 Smigielski EM, Sirotkin K, Ward M, *et al.* dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 2000;28:352–355.
31 Damm F, Heuser M, Morgan M, *et al.* Integrative prognostic risk score in acute myeloid leukemia with normal karyotype. Blood 2011;117:4561–4568.
32 Zhang Y, Strissel P, Strick R, *et al.* Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia. Proc Natl Acad Sci USA 2002;99:3070–3075.
33 Altschul SF, Madden TL, Schaffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
34 Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med 2010;2:84.
35 Tsai AG, Lieber MR. Mechanisms of chromosomal rearrangement in the human genome. BMC Genomics 2010;11(Suppl 1):S1.
36 Feldman AL, Dogan A, Smith DI, *et al.* Discovery of recurrent t(6;7) (p25.3;q3.23) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. Blood 2010;117:915–919.
37 Delhommeau F, Dupont S, Della Valle V, *et al.* Mutation in TET2 in myeloid cancers. N Engl J Med 2009;360:2289–2301.
38 Clark JP, Cooper CS. ETS gene fusions in prostate cancer. Nat Rev Urol 2009;6:429–439.
39 Soda M, Choi YL, Enomoto M, *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561–566.
40 Navin N, Kendall J, Troge J, *et al.* Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90–94.
41 Duncavage EJ, Magrini V, Becker N, *et al.* Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J Mol Diagn 2011;13:325–333.

Supplementary Information accompanies the paper on Modern Pathology website (http://www.nature.com/modpathol)