The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias

Anna K Andersson1,2,14, Jing Ma1,14, Jianmin Wang3, Xiang Chen3, Amanda Larson Gedman1, Jinjun Dang1, Joy Nakitandwe1, Linda Holmfeldt1, Matthew Parker3, John Easton4, Robert Huether3, Richard Kriwacki5, Michael Rusch3, Gang Wu3, Yongjin Li3, Heather Mulder4, Susana Raimondi1, Stanley Pounds5, Guolian Kang6, Lei Shi6, Jared Becksfort3, Pankaj Gupta3, Debbie Payne-Turner1, Bhavin Vadodaria4, Kristy Boggs4, Donald Yergeau4, Jayanthi Manne4, Guangchun Song1, Michael Edmonson3, Panduka Nagahawatte3, Lei Wei3, Cheng Cheng6, Deqing Pei6, Rosemary Sutton7, Nicola C Venn7, Albert Chetcuti8, Amanda Rush8, Daniel Catchpoole8, Jesper Heldrup9, Thoas Fioretos2, Charles Lu10,11, Li Ding10,11, Ching-Hon Pui1,12, Donald Yergeau4, Jayanthi Manne4, Guangchun Song1, Michael Edmonson3, Panduka Nagahawatte3, Lei Wei3, Cheng Cheng6, Deqing Pei6, Rosemary Sutton7, Nicola C Venn7, Albert Chetcuti8, Amanda Rush8, Daniel Catchpoole8, Jesper Heldrup9, Thoas Fioretos2, Charles Lu10,11, Li Ding10,11, Ching-Hon Pui1,12, Sheila Shurtleff3, Charles G Mullighan1, Elaine R Mardis10,11, Richard K Wilson10,11, Tanja A Gruber1,12, Jinghui Zhang3 & James R Downing1 for The St. Jude Children’s Research Hospital–Washington University Pediatric Cancer Genome Project13

Infant acute lymphoblastic leukemia (ALL) with MLL rearrangements (MLL-R) represents a distinct leukemia with a poor prognosis. To define its mutational landscape, we performed whole-genome, exome, RNA and targeted DNA sequencing on 65 infants (47 MLL-R and 18 non–MLL-R cases) and 20 older children (MLL-R cases) with leukemia. Our data show that infant MLL-R ALL has one of the lowest frequencies of somatic mutations of any sequenced cancer, with the predominant leukemic clone carrying a mean of 1.3 non-silent mutations. Despite this paucity of mutations, we detected activating mutations in kinase-P13K-RAS signaling pathway components in 47% of cases. Surprisingly, these mutations were often subclonal and were frequently lost at relapse. In contrast to infant cases, MLL-R leukemia in older children had more somatic mutations (mean of 6.5 mutations/case versus 1.3 mutations/case, P = 7.15 × 10−5) and had frequent mutations (45%) in epigenetic regulators, a category of genes that, with the exception of MLL, was rarely mutated in infant MLL-R ALL.

ALL arising in infants less than 1 year of age accounts for 2.5–5% of all childhood ALL. Up to 80% of infant ALL cases are characterized by rearrangements of the MLL gene (encoding mixed-lineage leukemia) at 11q23 (also called KMT2A)1,2. Although current event-free survival rates for childhood ALL have reached greater than 85% (ref. 3), the outcome for infant ALL with MLL-R remains poor, with an event-free survival of only 28–36% (refs. 1–3). Thus, new therapeutic approaches are needed to improve cure rates for these patients. Studies on this leukemia subtype have demonstrated that the MLL translocation often occurs in utero and that clinically overt leukemia develops with a very short latency, with rare cases presenting at birth4,5. These observations suggest that infant MLL-R ALL requires few additional mutations to induce full transformation. Consistent with this hypothesis, genome-wide studies on MLL-R ALL using SNP arrays have shown that this leukemia subtype contains on average only one copy number alteration (CNA) per case6.7. Nevertheless, mouse models of MLL-R leukemia suggest that expression of the MLL chimeric gene alone is insufficient for full transformation8. To gain a better understanding of the complete landscape of somatic mutations in infant MLL-R ALL, we performed a genome-wide analysis on this leukemia subtype as part of the St. Jude Children’s Research Hospital–Washington University Pediatric Cancer Genome Project (PCGP)9.

RESULTS
Infant MLL-R ALL has an exceedingly low mutation frequency
We performed paired-end whole-genome sequencing on diagnostic leukemia cells and matched remission bone marrow or peripheral
blood cells from a discovery cohort of 22 infants with MLL-R ALL (Supplementary Fig. 1, and Supplementary Tables 1 and 2). The leukemic genomes had an average haploid coverage of 39x. Somatic alterations, including single-nucleotide variations (SNVs), insertions-deletions (indels), structural variations and CNAs, were detected using multiple analytical pipelines\textsuperscript{10,11} and validated using orthogonal DNA sequencing approaches (Online Methods).

Whole-genome sequencing showed that the infant MLL-R ALL genomes contained an average of 111 somatic sequence mutations (range of 29–229) and 10 CNAs or structural variations (range of 2–26) per case (Supplementary Fig. 2 and Supplementary Tables 3 and 4), yielding a median somatic coding mutation rate of \(7.30 \times 10^{-8}\) mutations per base (range of 0 to \(2.29 \times 10^{-7}\)). With the exception of pediatric low-grade glioma\textsuperscript{12}, this rate is 2- to 180-fold lower than the rates reported in adult and pediatric cancers (Supplementary Table 5). The mutation spectrum across the cohort did not suggest any specific mutational mechanisms, with the most common nucleotide changes being C>T/G>A transitions (Supplementary Fig. 2). The number of somatic alterations affecting the coding region of annotated genes or regulatory RNAs averaged 8.2 alterations/case, including 2.2 non-silent SNVs (range of 0–4) and 6.0 CNAs or structural variations (range of 2–19) per case (Fig. 1a–c, Supplementary Figs. 2 and 3, and Supplementary Tables 3, 4 and 6–8). RNA sequencing (RNA-seq) in 21 of 22 cases demonstrated that 48% of the non-silent SNVs were expressed (Supplementary Figs. 4 and 5, and Supplementary Tables 9–11). Despite the paucity of mutations, 81% of the expressed missense mutations were predicted to have a deleterious effect on protein function.

**Figure 1** Somatic mutations detected by whole-genome sequencing of infant MLL-R ALL. (a–c) Circos plots of somatic non-silent mutations in three infant MLL-R ALL cases. (a) INF006 contained a balanced t(4;11)(q21;q23) encoding the MLL-AFF1 fusion and gain of chromosome X. (b) INF009 contained a balanced t(11;19)(p13.3;q23) with the break on chromosome 19 occurring 22.7 kb 5′ of MLLT1, resulting in a spliced (as indicated by the asterisk) in-frame MLL-MLLT1 chimeric gene. The translocation had an inverted intrachromosomal duplication spanning 0.3 kb at the breakpoint on chromosome 19 that resulted in an out-of-frame RFX2-MLL fusion. In addition, an unrelated intrachromosomal deletion of 46 bp was detected on chromosome 6. (c) INF004 contained a complex three-way translocation involving chromosomes 9, 10 and 11 that encoded an MLL-MLLT10 fusion gene on the derivative chromosome 11, an out-of-frame MLLT10-CNTNAP3B chimeric gene on chromosome 9 and 3′-truncated MLL on chromosome 10. This case also contained a t(8;14)(q24;q11.2) that resulted in the juxtaposition of MYC with the TRA locus (encoding T cell receptor α), as denoted by the caret. In addition, a non-silent SNV in COL13A1 and a deletion on chromosome 9 disrupted PAX5. LOH, loss of heterozygosity; SV, structural variation. (d) Structure of the INF004 complex translocation involving chromosomes 9p11 (blue), 10p12 (yellow) and 11q14–23 (burgundy). Genomic coordinates (hg18) are indicated above each genomic segment. The MLL-MLLT10 fusion gene is depicted by the solid arrow; all other rearrangements are depicted by dotted arrows. Copy number gains (red) and losses (blue) at the respective breakpoints are shown. The final genomic products on chromosomes 9, 10 and 11 are shown in Supplementary Figure 9. Ter, terminus; Cen, centromere.
Figure 2 Recurrently mutated genes detected in 47 cases of infant MLL-R ALL. Whole-genome sequencing was performed on 22 leukemic samples (discovery cohort), and targeted capture and sequencing of the coding exons of 232 genes was performed on an additional 25 infant MLL-R ALL samples (validation cohort). (a) Recurrent mutations were identified in 21 genes or loci in the combined infant MLL-R ALL cohorts. An asterisk indicates that more than two genes were targeted within the locus (Supplementary Tables 7, 8, 14, 19 and 24). (b) Distribution of mutated genes across the 47 infant MLL-R cases. For mutations in kinase genes or in genes in the PI3K-RAS pathway, only those known to confer activation of the pathway are shown. An asterisk indicates that more than two genes were targeted within the locus in one or more of the cases, with only the first gene in the locus listed. A dagger indicates that the sample lacked a matched non-leukemic sample. A caret indicates a sample for the identical twin of sample INF060; the matched non-leukemic sample from INF060 was also used for the twin. A number sign indicates a non-leukemic sample from INF060 was also used for the twin. An open circle designates that the mutant allele is expressed, as determined by RNA-seq, and a dash indicates that it is not expressed at the level of detection for our analysis.

Notably, the majority of the CNAs and structural variations were a direct consequence of the MLL rearrangement, with the breakpoint of both MLL and its partner gene typically accompanied by the gain or loss of adjacent genetic material (Fig. 1d). When the MLL-related CNAs and structural variations were removed from consideration, the average number of CNAs and structural variations per case decreased to 3.0 (Supplementary Table 12). Additionally, we detected 40% of the identified somatic SNVs at a mutant allele frequency (MAF) below 30% despite the high tumor purity (average of 92%, range of 74–100%; Supplementary Tables 1, 4, 8 and 13). This suggests that these mutations reside in minor diagnostic subclones. We confirmed this hypothesis by performing custom capture and deep sequencing at an average coverage of 437x for all the tier 1–3 somatic mutations identified in five cases. This analysis demonstrated the presence of substantial intratumoral heterogeneity (Supplementary Figs. 6 and 7). If we focused only on the mutations (including SNVs and indels) present in the dominant leukemic clone, the average number of non-silent mutations decreased to 1.3 mutations/case (Supplementary Table 13), with only 0.6 expressed at the RNA level (Supplementary Fig. 5 and Supplementary Table 11). A direct comparison of these mutation frequencies to those observed across 29 human cancers showed infant ALL to have one of the lowest frequencies of somatic mutation (Supplementary Fig. 8)13–15.

The majority of MLL rearrangements are unbalanced

Whole-genome sequencing detected a total of 133 structural variations and CNAs across the infant ALL cohort affecting annotated genes (Supplementary Tables 3, 7, 12 and 14, and Supplementary Note). Approximately half (n = 67) of the 133 structural variations and CNAs were a direct result of the MLL rearrangement. Within the remaining half (n = 66), the most frequently affected genes were PAX5 (five cases), CDK2NA/CDK2NB (three cases) and the noncoding RNA genes DLEU1/DLEU2 (three cases). The structural variations and CNAs affecting these genes were predicted to result in loss of function of a single allele.

Consistent with previous studies16–19, more than half of the MLL rearrangements were complex, involving three or more chromosomes and/or accompanied by large insertions, deletions and/or inversions of sequences adjacent to the breakpoints (Fig. 1d, Supplementary Figs. 3 and 9–16, and Supplementary Table 8). Moreover, even so-called simple cytogenetically balanced MLL translocations that involved only two chromosomes were found at the base-pair level to have focal deletions and/or insertions of sequences at the breakpoints (Supplementary Table 15). As a result, although each MLL rearrangement was predicted to encode an in-frame MLL–partner gene fusion protein, only 10 of 22 of the analyzed cases were predicted to encode an in-frame reciprocal partner gene–MLL fusion protein. RNA-seq on available samples demonstrated that six of nine cases with a predicted reciprocal fusion expressed the reciprocal product, consistent with previous reports (Supplementary Tables 16 and 17)20. Two of the predicted in-frame reciprocal fusion proteins involved genes with a known role in cancer: KRAS-MLL and AFF1-RAD51B-MLL (Supplementary Figs. 11 and 12, and Supplementary Table 17). Some of the complex MLL rearrangements also resulted in alterations of genes adjacent to MLL and/or the MLL partner gene (Supplementary Table 16). An analysis
of the sequence surrounding the breakpoints of  MLL and its partner genes suggested that the predominant mechanism of rearrangement involved non-homologous end joining\textsuperscript{21}.

We performed RNA-seq on an additional 12 diagnostic MLL-R cases from a validation cohort of infant MLL-R ALL and identified a new MLL fusion partner, USP2 (encoding ubiquitin-specific peptidase 2), located at 11q23.3, approximately 1 Mb from MLL on the reverse strand (Supplementary Fig. 17 and Supplementary Table 17). RNA-seq also identified two new non-MLL in-frame fusion genes in INF016, CABIN1-TRAPPC10 and PAX5-KANK1 (Supplementary Table 17), as well as an out-of-frame DDTL-CABIN1 fusion. Upon manual review, we identified CABIN1-TRAPPC10 and DDTL-CABIN1 fusions in very few whole-genome sequencing reads, whereas the PAX5-KANK1 fusion lacked any whole-genome sequencing reads, suggesting that these fusions were present in minor subclones.

Mutations in tyrosine kinase–PI3K-RAS signaling pathways

Despite the paucity of somatic mutations in the discovery cohort, we observed activating mutations in tyrosine kinase–phosphoinositide 3-kinase (PI3K)-RAS pathways, with recurrent mutations in KRAS (n = 4) and NRAS (n = 2) and non-recurrent mutations in FLT3, NFI, PTPN11 and PIK3R1 (Supplementary Tables 6 and 8). In contrast to the non-silent SNVs, where only 48% of the mutant alleles were expressed, 100% of the activating kinase–PI3K-RAS pathway mutant alleles were expressed (Supplementary Fig. 5 and Supplementary Table 11). To extend these results, we sequenced the exons of 232 genes that included all mutated genes identified in the discovery cohort, as well as other genes in kinase–PI3K-RAS signaling pathways (for a list of the sequenced genes, see the Supplementary Note), in a validation cohort consisting of an additional 43 infant ALL cases, of which 25 harbored an MLL rearrangement. Each sample was also analyzed for CNAs by SNP arrays (Online Methods and Supplementary Tables 18–23). Recurrent mutations were identified in 21 genes or gene loci across the combined infant MLL-R ALL cohorts (Fig. 2a and Supplementary Table 24). Notably, we identified activating mutations in tyrosine kinase–PI3K-RAS pathway genes in 22 of 47 (47%) of the infant MLL-R cases (Fig. 2a,b, Supplementary Figs. 18 and 19, and Supplementary Table 25). The tyrosine kinase–PI3K-RAS pathway mutations were observed in association with each of the different types of MLL rearrangements identified in infant ALL (Fig. 2b). In every case analyzed by RNA-seq, the activating mutant alleles were expressed, irrespective of their MAFs (Fig. 2b and Supplementary Table 11). Furthermore, gene set enrichment analysis within the cohort positive for the MLL-AFF1 fusion identified the presence of expression signatures consistent with RAS pathway activation (Supplementary Figs. 20–23 and Supplementary Table 26).

The majority of the identified tyrosine kinase–PI3K-RAS pathway mutations have previously been shown to be activating mutations (or inactivating mutations in the case of the homozygous deletion of NF1 in INF018)\textsuperscript{22–25}. However, two cases contained a FLT3 mutation encoding p.Asn676Lys that was previously reported in an acute myeloid leukemia (AML) case after the development of resistance to the kinase inhibitor PKC412 (refs. 30,31) and as a recurrent mutation in core binding factor leukemia\textsuperscript{32} (Supplementary Fig. 24 and Supplementary Note). Similarly, we identified two cases that contained new PIK3R1 mutations, including an internal tandem duplication in the inter-SH2 domain (PIK3R1-ITD) (Supplementary Figs. 25–27 and Supplementary Note). We functionally demonstrated that the FLT3 (encoding p.Asn676Lys) and PIK3R1 mutations were activating, resulting in factor-independent growth of the interleukin (IL)-3–dependent mouse leukemia cell line BaF3 (Supplementary Figs. 28 and 29).
A surprising observation was that 65% (20/31, observed in 22 cases) of the activating tyrosine kinase–PI3K-RAS pathway mutations had MAFs <30%, suggesting that they were present in minor subclones (Supplementary Tables 25 and 27, and Supplementary Note). Moreover, although eight cases contained two or more mutations in this pathway, the MAF for each mutation was below 30% in six cases. To further explore the importance of the identified tyrosine kinase–PI3K-RAS pathway mutations, we analyzed seven matched diagnostic and relapse infant MLL-R samples, including five with a mutation in this pathway at diagnosis. Although the MAF of the activating mutation was maintained or increased at relapse in two cases (INF001 and INF65), the mutations were lost in two other cases (INF033 and INF042) and, in another case (INF073), the MAF for the PIK3CA mutation decreased from 39% to 15% (Supplementary Table 28). These data suggest that activating mutations in tyrosine kinase–PI3K-RAS pathways may not be necessary for the maintenance of leukemic cells.

Although we observed a trend toward poorer event-free survival and overall survival and an increased cumulative incidence of relapse in patients harboring an activating pathway mutation, these trends did not reach statistical significance (Supplementary Fig. 30). There was no difference in the age at diagnosis of infant MLL-R cases with or without an activating mutation (Supplementary Fig. 31). However, within the infant cohort positive for the MLL-AFF1 fusion (n = 23), those with an activating mutation were on average younger at diagnosis than those lacking a mutation (mean of 4.01 versus 6.75 months, \( P = 0.0473 \); Supplementary Fig. 32a). Furthermore, there was an age difference among patients with activating mutations in a major or minor clone and those lacking an activating mutation (means of 2.77, 5.09 and 6.75 months, respectively, \( P = 0.0228 \); Supplementary Fig. 32b), suggesting that an activating mutation among MLL-AFF1–positive patients may be associated with decreased disease latency.

Clonal evolution at relapse in infant MLL-R ALL

To further explore the relationship between diagnosis and relapse in infant MLL-R ALL, we compared all somatic mutations identified in the diagnostic and relapse samples for two infant patients with MLL-R ALL, each relapsing 3 years after diagnosis (INF001D/INF001R and INF002D/INF002R). We analyzed non-tumor as well as diagnostic and relapse samples by whole-genome sequencing followed by custom capture enrichment and deep sequencing, allowing us to accurately calculate the MAFs for all SNVs in each leukemia. In both cases, similarly to non–MLL-R leukemia, relapse was associated with a marked increase in the total number of SNVs and CNAs or structural variations (Fig. 3a,b, Supplementary Figs. 33–35, and Supplementary Tables 3, 29 and 30).

We analyzed the SNVs occurring in unique sequences outside of genes or gene regulatory regions (tier 3 SNVs) at diagnosis and relapse for each case by assigning them to clusters with the highest probability on the basis of their MAFs (Online Methods). We identified five mutation clusters in INF002 at diagnosis with MAFs of 0.426, 0.236, 0.114, 0.058 and 0.024 (Fig. 3c,d). The diagnostic mutations with a MAF of 0.426 are consistent with heterozygous mutations present in every leukemic cell, whereas clusters with lower MAFs represent mutations present in minor subclones. After adjusting for tumor purity, the diagnostic sample was predicted to contain five related clones with 6.25–50% population frequencies. At relapse, the majority of tier 3 mutations occurred at a MAF of 0.5, suggesting the presence of a single clone descending from a minor clone present as 6.25% of the diagnostic sample. The results for INF001 followed the same trend (Supplementary Fig. 33). In both cases, the founder clone at relapse acquired additional mutations after treatment, although the mutation spectrum did not differ substantially between the diagnostic and relapse samples for these two patients (Supplementary Fig. 36).
Genomic landscape of infant non–MLL-R ALL
We also performed targeted gene resequencing of the 232 genes and SNP arrays on 18 infant non–MLL-R ALLs (Online Methods and Supplementary Tables 18 and 21–23). There was no significant difference in the average number of SNVs in these 232 genes or in the frequency of tyrosine kinase–PI3K-RAS pathway mutations between infant MLL-R (22/47, 47%) and infant non–MLL-R (7/18, 39%) ALL (P = 0.59); again, all activating mutations were expressed in the 6 cases subjected to RNA-seq. In contrast to SNVs, CNAs were found significantly more often in infants lacking an MLL rearrangement than in infants with MLL rearrangements (average of 2.2/case versus 1.0/case); however, this increased frequency was associated with age and was independent of the presence or absence of the MLL rearrangement (P = 0.005; Supplementary Figs. 37–39). CDKN2A/CDKN2B deletions were more frequent in the infant non–MLL-R cases, approaching the frequency seen in standard non–Philadelphia-positive childhood ALL (5/18, 28% versus 4/47, 8.5%, P = 0.058)33.

Six of the 18 non–MLL-R infant cases had RNA available for sequencing, 4 of which carried either newly identified fusion genes and/or altered genes that would be predicted to encode truncated proteins (Supplementary Table 17). Specifically, we identified two new fusions containing NUTM1: BRD9-NUTM1 (INF049) and ACIN1-NUTM1 (INF074). The BRD9-NUTM1 fusion product created by t(5;15)(p15;q14) consisted of the first 14 exons of BRD9 fused to exons 3–8 of NUTM1. This newly identified fusion is analogous to the BRD3-NUTM1 and BRD4-NUTM1 fusions found in midline carcinomas34. The patient we analyzed was described as having a t(5;15)(p15;q12), which is a previously reported cytogenetic finding in infant non–MLL-R ALL; the BRD9-NUTM1 fusion may thus be recurrent in infants with t(5;15)(p15;q11-13)35,36. The ACIN1-NUTM1 fusion involved the same exons of NUTM1 fused to the first four exons of ACIN1 at 14q11. The patient we analyzed had an ins(15;14)(q22;q11.2q32.1), by conventional karyotyping, indicating that NUTM1 may be a candidate gene for cytogenetic rearrangements involving 15q (ref. 37).

INF061 had both a BICD2-JAK2 and an ARHGAP32-CAPRIN1 fusion, as well as a truncation of the cohesin gene STAG2 (Supplementary Table 17). Finally, INF070 carried a truncation in ARID1B, a member of the SWI/SNF chromatin-remodeling complex; truncations of this gene have been described in 7% of pediatric neuroblastoma38. Overall, the structural variations and CNAs differed substantially between MLL-R and non–MLL-R infant ALL, underscoring their differences in biology and clinical outcome1.

Epigenetic mutations in non-infant MLL-R leukemia
Although MLL rearrangements occur at a high frequency in infant ALL, such genetic lesions are also seen in older children with ALL or AML39. To compare the MLL-R mutational profiles for infants and older children, we performed whole-exome sequencing and SNP array analysis on 20 non-infant patients with MLL-R leukemia (7–19 years of age; 9 ALLs, 10 AMLs and 1 case of acute undifferentiated leukemia), as well as RNA-seq on 18 of 20 cases (Supplementary Figs. 4 and 40, and Supplementary Tables 9–11, 17, 31, and 32). This analysis showed that the major clone in non-infant MLL-R leukemia harbored a significantly higher number of non-silent somatic SNVs and indels than in infant MLL-R ALL (mean of 6.5/case versus 1.3/case, P = 7.15 × 10−5; for expressed genes, mean of 3.2/case versus 0.6/case, P = 1.6 × 10−3; Fig. 4a and Supplementary Tables 33 and 34; see Supplementary Table 35 for a mutation summary for all three cohorts). Although there was a trend toward a lower basal mutation rate in infants in comparison to non-infants (P = 0.15), multiple linear regression analysis demonstrated that the significantly higher number of mutations in older children could not be attributed solely to the difference in the basal mutation rate (Supplementary Figs. 41 and 42, and Supplementary Note). This finding suggests that overt leukemia in older children with MLL rearrangements may require the acquisition of more cooperating mutations. As in infants with an MLL rearrangement, we identified activating mutations in tyrosine kinase–PI3K-RAS pathway genes in 50% of the non-infant leukemias, with recurrent mutations in FLT3 (n = 3), KRAS (n = 3) and NRAS (n = 3) and non-recurrent mutations in CBL, PIK3CD, PTPN11 and PPM1I, all of which were expressed at the RNA level (Supplementary Figs. 43 and 44, and Supplementary Tables 11 and 36). In contrast to the mutation distribution in infant MLL-R ALL cases, the majority of the mutations in tyrosine kinase–PI3K-RAS pathway genes in the non-infant MLL-R leukemias were present in the major clone (Supplementary Fig. 44 and Supplementary Table 25). This observation extended to all identified somatic exonic mutations, whereas, in infant MLL-R ALLs, these mutations were more commonly seen in minor clones (P < 0.0001; Supplementary Fig. 45). Non-infant MLL-R leukemias had a significantly higher number of CNAs as compared to infant MLL-R ALLs (average of 2.6/case versus 1.0/case, P = 0.0234; Supplementary Fig. 46). We noted deletions involving CDKN2A/CDKN2B in three of nine ALL cases (33%) but in none of the AML cases. None of the 20 non-infant MLL-R leukemias harbored a focal PAX5 lesion, which is in contrast to infant MLL-R ALLs, where PAX5 alterations were present in 5 of 22 cases (23%) (Supplementary Figs. 39, 44 and 47, and Supplementary Tables 37–39). RNA-seq identified two new in-frame non-MLL fusions: SETD2-CCDC12 (SJMLL009) and PABPC1L–YWHAB (SJMLL019). In addition, eight events identified in six cases resulted in out-of-frame fusions, one of which included MLL and four of which included MLL partner genes (Supplementary Table 17). At the RNA level, we noted a unique gene expression signature correlating with age when comparing infants to older children with the MLL-AFFI1 fusion (Supplementary Fig. 48, and Supplementary Tables 40 and 41).

An interesting observation in non-infant MLL-R leukemias was the presence of somatic mutations in genes whose products have a direct role in epigenetic regulation. The dominant epigenetic regulatory proteins are encoded by 633 genes (Supplementary Table 42)40. If we excluded MLL, which was altered in every case, we identified somatic mutations in 11 epigenetic regulatory genes (CHD4, SETD2, CREBBP, L3MBTL3, ATR, KAT6A, KDM6A, NSD1, PARP8, SUPT3H1 and TET3) in 9 of 20 (45%) of the non-infant MLL-R leukemias (Fig. 4b and Supplementary Fig. 44). By contrast, only 3 of 22 (14%) infant MLL-R ALL cases harbored somatic mutations in epigenetic regulatory genes (P = 0.04; Fig. 4b).

DISCUSSION
In this analysis of the genomic landscape of infant MLL-R ALL, we demonstrate that this highly aggressive leukemia contains remarkably few somatic mutations, having one of the lowest somatic coding mutation rates observed in a human cancer thus far41,42. The only sequenced cancer that has a lower number of somatic coding mutations is pediatric low-grade glioma, a clinically indolent tumor12. The observed low mutational burden is consistent with the known oncogenic potency of MLL fusion proteins and the very short disease latency seen in infant MLL-R ALL, with some patients presenting with overt leukemia at birth. However, despite the low overall number of cooperating mutations, tyrosine kinase–PI3K-RAS signaling pathways were activated in almost half of infant MLL-R ALL cases. Although mutations in RAS genes have previously been described in MLL-R ALL43–45, we demonstrate for the first time, to our knowledge, recurrent activating
mutations targeting the PI3K complex. Specifically, we found activating mutations in PIK3CA and PIK3R1 in 11% of the cases. The high frequency of activating mutations in tyrosine kinase–PI3K–RAS pathway genes underscores the biological cooperativity between MLL fusion proteins and enhanced signaling through these pathways46-48.

Although the allelic frequencies of the MLL chimeric genes indicated that these rearrangements are present in every leukemic cell, our detailed analysis of MAFs for the other identified somatic mutations suggest that they are often present in minor clones. For example, 65% of the activating tyrosine kinase–PI3K–RAS mutations had MAFs <30%, suggesting that they were present in minor diagnostic clones. In addition, although eight infant MLL–R ALL cases harbored two or more mutations in genes in these pathways, the MAF for each mutation was below 30% in six of these cases. Notably, all of the activating tyrosine kinase–PI3K–RAS pathway mutations were expressed, which was in contrast to other non-silent mutations in the discovery cohort, of which only 38% were expressed. Even more surprisingly, in our analysis of seven matched diagnostic and relapse infant MLL–R leukemia samples, five of which contained a diagnostic mutation in tyrosine kinase–PI3K–RAS pathways, we found that these mutant genes were maintained at relapse in two cases, decreased in frequency in one case and lost in two cases. Consistent with these observations, a recent report describing three paired samples from cases with MLL-AFF1 fusion demonstrated that two lost the clone carrying the RAS gene mutation at relapse49.

The presence of activating mutations in RAS genes in infant MLL–R leukemia has previously been suggested to be an independent predictor of a poor prognosis 43. Although our cohort is small, our analysis suggested a trend toward a lower event-free survival and overall survival rate in cases with tyrosine kinase–PI3K–RAS pathway mutations, especially when the mutation was present at a MAF >30%. In contrast to the Interfant study 43, within the cohort positive for the MLL–AFF1 fusion, those with an activating mutation were on average younger than those lacking a mutation. This age difference was even more pronounced when considering the cases positive for the MLL–AFF1 fusion with an activating mutation in the major clone, indicating that the presence of an activating mutation in these pathways may decrease the time required for the development of clinically overt disease. These data suggest that the coexistence of the MLL fusion protein with an activating mutation in tyrosine kinase–PI3K–RAS signaling pathways is not essential for either the establishment or maintenance of leukemia; however, the presence of these alterations likely confers a growth advantage. Our data also raise the possibility that the high degree of clonal heterogeneity observed in infant MLL–R ALL 59 might directly contribute to its poor prognosis, in that conventional ALL treatment regimens may not fully eliminate all clones, allowing the emergence of relapsed disease from a minor clone present at the time of diagnosis.

Interestingly, in non-infant MLL–R leukemias, most somatic mutations were found to reside within a single dominant clone. In addition, these cases harbored significantly more non-silent mutations than infant MLL–R leukemias, suggesting that MLL fusion genes require an increased number of cooperating mutations in older patients to generate overt disease. This raises the possibility that there may be a fundamental difference in the target cell of transformation, the microenvironment or both between infant and older patients. This hypothesis is further supported by the finding that non-infant MLL–R leukemias carry significantly more mutations in genes encoding epigenetic regulators than infant cases, suggesting that the target cell in infants may already have a chromatin state that is more permissive to transformation by the MLL gene rearrangement. These data provide a unique look into the underlying molecular pathology of this highly aggressive pediatric leukemia. Our results suggest that, in infants, leukemia is initiated in a hematopoietic progenitor cell (HPC) that may have a different chromatin state than more mature HPCs and that, within this context, the MLL chimeric gene has a dominant role in establishing overtly leukemic cells. These cells gain few additional mutations during their proliferative expansion, although the frequent targeting of genes within tyrosine kinase–PI3K–RAS signaling pathways suggests that activation of this signaling can cooperate in leukemogenesis, albeit not to a level that results in the establishment of a dominant clone. Moreover, loss of the latter mutations at relapse in a number of patients suggests that specifically targeting these activating lesions will likely provide little therapeutic benefit. In contrast, the clear implication that the MLL fusion protein is a potent driver in this aggressive leukemia and that leukemic cells lack a high mutation rate raises the possibility that therapy directed toward the MLL fusion protein, or proteins required for its biological actions, might be a fruitful approach for targeted therapy in this aggressive leukemia.

URLs. PyMOL, http://www.pymol.org/; R statistical software, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The sequence and SNP microarray data have been deposited in the European Genome-phenome Archive (EGA) under study accession EGAS00001000246.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all the patients and their parents from the St. Jude Children’s Research Hospital (USA), the Department of Pediatrics at Lund University Hospital (Sweden) and the children’s hospitals associated with the Australian and New Zealand Children’s Haematology and Oncology Group. We thank B. Pappas and S. Malone for IT infrastructure. We thank the Tissue Resources Laboratory, the Flow Cytometry and Cell Sorting Core and the Clinical Applications of Core Technology Laboratories of the Hartwell Center for Bioinformatics and Biotechnology of St. Jude Children’s Research Hospital. We thank E. Parganas and J. Ihele (St. Jude Children’s Research Hospital) for the mouse BaF3 cells. This work was funded by the St. Jude Children’s Research Hospital–Washington University Pediatric Cancer Genome Project and the American Lebanese and Syrian Associated Charities of St. Jude Children’s Research Hospital and was supported by a grant from the US National Institutes of Health (P30 CA021765). A.K.A. was supported by the Swedish Childhood Cancer Society, the Swedish Research Council, the Swedish Cancer Society, BioCARE and the Gunnar Nilsson Cancer Foundation. C.G.M. is a Pew Scholar in Biomedical Sciences and a St. Baldrick’s Scholar.

AUTHOR CONTRIBUTIONS

A.K.A., J.R.D., T.A.G., J.Z., J. Ma and R.K.W. designed all experiments. J.Z. and J. Ma led the sequencing analysis. J.Z., J. Ma, J.W., X.C., M.P., M.R., G.W., Y.L., J.B., P.G., M.E., P.N., L.W., C.L., I.D. and E.R.M. performed the computational data analyses. G.S. provided bioinformatics support. A.K.A., L.H. and C.G.M. analyzed SNP array data. R.H. and R.K. performed structural modeling. A.L.G. and J.D. performed functional work on the PI3K and PIK3 pathway mutations. J.N., J.E., M.P., B.V. and D.Y. performed validation experiments. K.B. performed RNA-seq. J.E. and J. Manne performed exome sequencing. C.G.M., H.M. and D.P.-T. prepared samples. S.P., G.K., L.S., C.C. and D.P. performed statistical analysis. R.S., N.C.V., A.C., A.R., D.C., J.H., T.F. and C.-H.P. provided annotated patient samples. S.R. and S.S. provided molecular genetics, cytogenetics and FISH data. J. Ma and T.A.G. performed critical reading and contributed to the writing of the manuscript. A.K.A. and J.R.D. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.
1. Pieters, R. et al. Treatment protocol for infants younger than 1 year with acute lymphoblastic leukemia (Interfant-99): an observational study and a multicentre randomised trial. Lancet Oncol. 14, 703–711 (2013).

2. Bondi, A., Cimino, G., Pieters, R. & Pui, C.H. Biological and therapeutic aspects of infant leukemia. Blood 96, 24–33 (2000).

3. Pui, C.H. et al. Treating childhood acute lymphoblastic leukemia without cranial irradiation. N. Engl. J. Med. 360, 2730–2741 (2009).

4. Field, A.M. et al. In utero rearrangements in the trithorax-related oncogene in infant leukemias. Nature 363, 358–360 (1993).

5. Gale, K.B. et al. Backtracking leukemia to birth: identification of clonotopic gene fusion sequences in neonatal blood spots. Proc. Natl. Acad. Sci. USA 94, 13950–13954 (1997).

6. Mullighan, C.G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukemia. Nature 446, 758–764 (2007).

7. Bardini, M. et al. Implementation of array based whole-genome high-resolution technologies confirms the absence of secondary copy-number alterations in ALL-AF4-positive infant ALL patients. Leukemia 25, 175–178 (2011).

8. Krivtsov, A.V. & Armstrong, S.A. MLL translocations, histone modifications and leukemia stem-cell development. Nat. Rev. Cancer 7, 693–703 (2010).

9. Downing, J.R. et al. Identifying ALL patients with a poor outcome in remission by karyotypic analyses of recurrent genomic translocations. Blood 113, 2432–2434 (2009).

10. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods 8, 652–654 (2011).

11. Zhang, J. et al. The genetic basis of early T-cell precursor acute lymphoblastic leukemia. Nature 481, 157–163 (2012).

12. Zhang, J. et al. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. Nat. Genet. 45, 602–612 (2013).

13. Loh, M.L. et al. MLL rearrangements: in vivo and in vitro in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).

14. Wu, G. et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nat. Genet. 44, 251–253 (2012).

15. Zhang, J. et al. A novel retinoblastoma oncogene therapy from genomic and epigenetic analyses. Nature 481, 329–334 (2012).

16. Gillert, E. et al. DNA damage repair mechanism is involved in the origin of chromosomal translocations t(4;11) in primary leukemic cells. Oncogene 18, 4663–4671 (1999).

17. Reichel, M. et al. Fine structure of translocation breakpoints in leukemia blasts with chromosomal translocation t(4;11): the DNA damage-repair model of leukemogenesis. Oncogene 17, 3035–3044 (1998).

18. Super, H.O. et al. Identification of complex genomic breakpoint junctions in the t(9;11) MLL-AF4 fusion gene in acute leukemia. Genes Chromosom. Cancer 20, 185–195 (1997).

19. Meyer, C. et al. The MLL recombine in acute leukemias in 2013. Leukemia 27, 2165–2176 (2013).

20. Downing, J.R. et al. The der(11)-encoded MLL-AF4 fusion transcript is consistently detected in t(4;11)(q21;q23)-containing acute lymphoblastic leukemia. Blood 83, 330–335 (1994).

21. Zhang, Y. & Rowley, J.D. Chromatin structural elements and chromosomal translocations in leukemia. DNA Repair (Amst.) 5, 1282–1297 (2006).

22. Schubbert, S., Shannon, K. & Bollag, G. Hyperactive Ras in developmental disorders and cancer. Nat. Rev. Cancer 7, 295–308 (2007).

23. Tyner, J.W. et al. High-throughput sequencing screen reveals novel, transforming RAS mutations in myeloid leukemia patients. Blood 113, 1749–1755 (2009).

24. Smith, G. et al. Activating K-Ras mutations outwith ‘hotspot’ codons in sporadic colorectal tumours—implications for personalised cancer therapy. Br. J. Cancer 102, 693–703 (2010).

25. Janakiraman, M. et al. Genomic and biological characterization of exon 4 K-RAS mutations in human cancer. Cancer Res. 70, 5901–5911 (2010).

26. Smith, C.C. et al. Activity of ponatinib against clinically-relevant AC220-resistant kinase domain mutants of FLT3-ITD. Blood 121, 3165–3171 (2013).

27. Bentires-Alj, M. et al. Activating mutations of the Noonan syndrome-associated SHP2 PTPN11 gene in human solid tumors and adult acute myelogenous leukemia. Cancer Res. 64, 8816–8820 (2004).

28. Loh, M.L. et al. Mutations in PTPN11 implicate the SHP-2 phosphatase in leukaemogenesis. Blood 103, 2325–2331 (2004).

29. Heidel, A. et al. Clinical resistance to the kinase inhibitor PKC412 in acute myeloid leukemia by mutation of Asn-676 in the FLT3 tyrosine kinase domain. Blood 107, 293–300 (2006).

30. von Bubnoff, N. et al. FMS-like tyrosine kinase 3—internal tandem duplication tyrosine kinase inhibitors display a nonoverlapping profile of resistance mutations in vitro. Cancer Res. 69, 3032–3041 (2009).

31. Opat, S. et al. Exome sequencing identifies recurring FLT3 N676K mutations in core-binding factor leukemia. Blood 122, 1761–1769 (2013).

32. Mullighan, C.G., Williams, R.T., Downing, J.R. & Sher, C.J. Failure of CDKN2A/B (INK4A/ARF)-mediated tumor suppression and resistance to targeted therapy in acute lymphoblastic leukemia induced by BCR-ABL. Genes Dev. 22, 1411–1415 (2008).

33. Filipakopoulos, P. & Knapp, S. Targeting bromodomains: epigenetic readers of lysine acetylation. Nat. Rev. Drug Discov. 13, 337–356 (2014).

34. Heerema, N.A. et al. Cyto genetic features of infants less than 12 months of age at diagnosis of acute lymphoblastic leukemia: impact of the 11q23 breakpoint on outcome: a report of the Childrens Cancer Group. Blood 83, 2274–2284 (1994).

35. De Lorenzo, P. et al. Cytogenetics and outcome of infants with acute lymphoblastic leukemia and absence of MLL rearrangements. Leukemia 28, 428–430 (2014).

36. Heerema, N.A. et al. Abnormalities of chromosome bands 15q13-15 in childhood acute lymphoblastic leukemia. Cancer 94, 1102–1110 (2002).

37. Biondi, A., Cimino, G., Pieters, R. & Pui, C.H. Biological and therapeutic aspects of infant leukemia. Lancet Oncol. 11, 880–889 (2010).

38. Huether, R. et al. The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nat. Commun. 5, 3630 (2014).

39. Dobbins, S.E. et al. The silent mutational landscape of infant MLL-AF4 pro-Acute lymphoblastic leukemia. Genes Chromosom. Cancer 52, 954–960 (2013).

40. Chang, Y.Y., Basso, G., Sakamoto, K.M. & Nelson, S.F. Identification of somatic and germline mutations using whole exome sequencing of congenital acute lymphoblastic leukemia. BMC Cancer 13, 55 (2013).

41. Driessen, E.M. et al. Frequencies and prognostic impact of RAS mutations in MLL-rearranged acute lymphoblastic leukemia in infants. Haematologica 98, 937–944 (2013).

42. Liang, D.C. et al. K-Ras mutations and N-Ras mutations in childhood acute leukemias with or without mixed-lineage leukemia gene rearrangements. Cancer 106, 950–956 (2006).

43. Prell, C., Bursen, A., Dingermann, T. & Marschalek, R. Secondary mutations in t(4;11) leukemia patients. Leukemia 27, 1425–1427 (2013).

44. Kim, W.I., Matsuoka, H. & Imai, M. N676K mutations in the activating mutations of the Noonan syndrome–associated RAS oncogene suppression induces apoptosis followed by more differentiated and less myelosuppressive disease upon relapse of acute myeloid leukemia. Blood 113, 1086–1096 (2009).

45. Ono, R. et al. Mixed-lineage-leukemia (MLL) fusion protein collaborates with Ras to induce acute leukemia through aberrant Hox expression and Raf activation. Leukemia 23, 2197–2209 (2009).

46. Tamai, H. et al. Activated K-Ras protein accelerates human MLL-AF4-induced leukemo-lymphomogenicity in a transgenic mouse model. Leukemia 25, 888–891 (2011).

47. Bardini, M. et al. Clonal variegation and dynamic competition of leukemia-initiating cells in infant acute lymphoblastic leukemia with MLL rearrangement. Leukemia 29, 38–50 (2015).
Subjects. Paired-end whole-genome sequencing on diagnostic leukemia blasts and matched germline samples was performed on a discovery cohort of 22 infants with MLL-R ALL and 2 matched relapse samples (INF001R and INF002R) using the Illumina sequencing platform. The infants were treated at St. Jude Children’s Research Hospital (Memphis, Tennessee, USA) and diagnosed during a period from 1992–2008. The cohort consisted of 10 cases with t(4;11) (MLL-AFFI (AF4)), 5 cases with t(11;19) (MLL-MLLT1 (ENL)), 4 cases with t(10;11) (MLL-MLLT10 (AF10)) and 3 cases with t(9;11) (MLL-MLLT3 (AF9)). 11q23 rearrangements were evaluated by FISH to confirm MLL gene rearrangements and/or RT-PCR for MLL-AFFI (AF4), MLL-MLLT3 (AF9), MLL-MLLT10 (AF10), MLL-MLT1 (ENL) and MLL-ELL fusions. Samples were cytogenetically analyzed and screened by RT-PCR for the presence of ETV6-RUNXI, TCF3-PBX1 and BCR-ABL fusions as part of routine clinical diagnostics. See Supplementary Table 1 for the clinical and genetic characteristics of the 22 infant MLL-R ALL cases.

The validation cohort consisted of 43 additional infant leukemia cases, 25 of which contained an MLL rearrangement (Supplementary Table 18). For 38 of the cases, a non-leukemic normal (germline) sample was available. INF059 and INF060 were identical twins, and a matched non-leukemic sample was only available for INF060; thus, this sample was also used as a germline sample for INF059.

In addition, five matched relapse infant MLL-R leukemia cases were investigated using either exome sequencing (SJINF039 and SJINF061). Whole-exome sequencing was performed on 20 cases of non-infant MLL-R leukemia (1 acute undifferentiated leukemia, 9 ALL cases and 10 AML cases) (see Supplementary Table 31 for the clinical and genetic characteristics of these leukemias) and their matched normal, non-leukemic DNA sample. All patients were treated at St. Jude Children’s Research Hospital. At diagnosis, 11q23 rearrangements were detected as described above.

Tumor and germline samples were obtained with informed consent using a protocol approved by the St. Jude Children’s Research Hospital institutional review board. The study was approved by the institutional review boards of St. Jude Children’s Research Hospital and Washington University.

Copy number analyses using Affymetrix SNP arrays. Of the 22 infant MLL-R ALL cases from the discovery cohort, 10 had copy number analyses performed using Affymetrix 500k SNP array as part of other studies and 12 had copy number data from the Affymetrix 6.0 SNP array data. In this study, we performed copy number analysis using the Affymetrix 6.0 SNP array for all samples in the validation cohort and for the non-infant MLL-R cases.

Illumina library construction for whole-genome sequencing. All methods for library construction and whole-genome DNA sequencing have been described previously.50,51 Detailed information regarding sequencing coverage is included in Supplementary Figure 1 and Supplementary Table 2.

Agilent liquid capture and library construction for whole-exome sequencing. Exon sequence was captured from 3 µg of genomic DNA using the Agilent SureSelect Human All Exon 50Mb kit (Agilent Technologies) following the manufacturer’s instructions. After elution, cDNA libraries were sequenced (paired-end 2 × 101 cycles) on Illumina sequencers.

RNA sequencing. RNA-seq was performed on 21 diagnostic and 2 paired relapse cases from the discovery cohort (not INF007), on 12 diagnostic and 4 relapse MLL-R infant ALLs and 6 non–MLL-R infant leukemias from the validation cohort, and on 18 of 20 cases from the non-infant MLL-R cohort (Supplementary Tables 9–11). Total RNA was extracted with TRIzol (Ambion) according to the manufacturer’s protocol. Total RNA (1 µg) was treated with DNase I (Ambion) at room temperature for 15 min, and phenol-chloroform-isomyl extraction and ethanol precipitation were performed. The integrity of the DNase I–treated RNA was analyzed on an Agilent 2100 Bioanalyzer (Agilent Technologies) before selection of poly(A) mRNA. Selection for poly(A) mRNA and subsequent cDNA synthesis were carried out using the Illumina TruSeq RNA sample preparation kit according to the manufacturer’s protocol. cDNA was fragmented (200-bp peak) with a Covaris E210 ultrasonicator before library preparation using the Illumina TruSeq RNA sample preparation kit according to the manufacturer’s protocol. Adapter-ligated fragments were PCR amplified for ten cycles. The quality and size of the final library preparation were analyzed on an Agilent 2100 Bioanalyzer. Sequencing was performed on the Illumina HiSeq 2500 platform in rapid mode to generate paired-end 100-cycle reads. Putative in-frame fusions were validated by RT-PCR followed by Sanger sequencing. Primer sequences are available upon request.

Analysis of whole-genome sequencing data. The methods employed for whole-genome sequence mapping, coverage and quality assessment, SNV and indel detection, tier annotation for sequence mutations, prediction of deleterious effects of missense mutations and identification of LOH have been described previously.11 Briefly, transcripts from Ensembl52 (build 54_36) and Genbank3,34 (build downloaded on 21 May 2009) were used for annotation. Sequence variants were classified into the following four tiers: (i) tier 1: coding synonymous, nonsynonymous, splice-site and noncoding RNA variants; (ii) tier 2: conserved variants (conservation score cutoff of greater than or equal to 500, based on either the phastConsElements28way table or the phastConsElements17way table from the UCSC Genome Browser) and variants in regulatory regions annotated by UCSC (regulatory annotations included are targetScanS, OREGAnno, tbsConsSites, vistaEnhancers, epochrome, firstEF, lTAFIValid, Poly(A), switchDbTs, encodeUviennaRnaZ, laminB1 and cpgIslandExt); (iii) tier 3: variants in non-repeat masked regions; and (iv) tier 4: the remaining SNVs.

Structural variations, including interchromosomal translocations (CTX), intrachromosomal translocations (ITX), inversions (INV), deletions (DEL) and insertions (INS), were analyzed by CREST59 and annotated as previously described.11 In addition, BreakDancer55 was run using default parameters for INF013. Paired tumor-normal bam files were used to identify putative somatic structural variations. All predicted transcripts were annotated, and transcripts that would lack a coding-sequence (CDS) start or stop site were filtered out. CDS lengths were then computed by taking the sum of the lengths of the coding portions of the exons containing the CDS start and stop sites and the lengths of all intermediate exons. In cases where both breakpoints of an event occurred inside exons, we reduced both exons to a single fused exon that was built by assembling the reads at the structural variation breakpoints and aligning the assembled contig back to the reference. If the fused exon lay between the CDS start and stop sites, the fused exon’s length was included in the CDS length calculation. In addition, if the predicted CDS length was a multiple of three, the fusion transcript was classified as being ‘in frame’. If the transcript was altered from the annotated CDS of one of the genes, it was classified as being ‘modified in frame’. In cases where the fusion occurred in the UTR, there could be modified transcripts with unmodified CDSs, so we explicitly checked for and flagged transcripts predicted to have modified in-frame CDSs.

CNAs were identified by evaluating the difference in read depth for each tumor and its matching normal DNA sample using the algorithm CONSERTING11. Confidence for a CNA segment boundary was determined using a series of criteria, including the length of flanking segments, the difference in CNA between neighboring segments, the presence of sequence gaps on the reference genome, the presence of structural variation breakpoints and any CNA in the matching germline sample. All CNAs were manually reviewed and compared with Affymetrix SNP 6.0 or 500K CNA results. Regions with LOH were identified from the high-quality SNPs. First, heterozygous SNPs with a MAF between 40 and 60% in the germline sample were used to estimate the LOH signal. For each heterozygous SNP, the LOH signal was calculated as the absolute MAF difference between the tumor and germline samples. Second, we used regression tree to perform the segmentation to infer CNAs, and adjacent segments were merged if the LOH signals between them were not statistically significant (Bonferroni corrected P value > 0.05, unpaired t-test).

Analysis of RNA sequencing data. For RNA-seq, paired-end sequencing was performed using the Illumina HiSeq platform with 100-bp read length. Paired-end reads from RNA-seq were aligned to the following four database files using the BWA (0.5.10) aligner: (i) the human GRCh37-lite reference sequence; (ii) RefSeq; (iii) a sequence file representing all possible combinations of
of non-sequential pairs in RefSeq exons; and (iv) the AceView database flat file downloaded from UCSC representing transcripts constructed from human ESTs. The mapping results from (ii) to (iv) were translated to human reference genome coordinates. In addition, they were aligned using STAR 2.3.0 to the human GRCh37-lite reference sequence without annotations. A bam file was constructed by selecting the best alignment among the five mappings. Poor-quality mappings were improved using SAM4, when possible, to generate the final bam file. Coverage was calculated using an in-house pipeline. Structural variation detection was carried out using CICERO, a novel algorithm that uses de novo assembly to identify structural variation in RNA-seq data (Y.L., T. Bo, M.R., J.E. and K.B. et al., unpublished data). Putative fusions were validated by RT-PCR.

Digital gene expression profiling. Transcript expression levels were estimated as fragments per kilobase of transcript per million mapped reads (FPKM); gene FPKMs were computed by summing the transcript FPKMs for each gene using Cuffdiff2 (refs. 56,57). A gene was considered ‘expressed’ if its FPKM value was 20.5 on the basis of the distribution of FPKM gene expression levels. Genes that were not expressed in any sample were excluded from the final data matrix for downstream analysis. RNA expression profiles were analyzed using Quicore Omics.

Targeted sequencing in the infant MLL-R and non–MLL-R validation cohorts. To determine the frequency of the identified mutations in a larger infant cohort, additional sequencing was performed for a total of 232 genes by Agilent Targeted Capture followed by Illumina sequencing (216 genes) or by PCR followed by Sanger sequencing (16 genes) in a validation cohort consisting of an additional 43 infant ALL cases, 25 of which harbored an MLL rearrangement. The 232 genes included in targeted sequencing were selected on the basis of the following criteria, including (i) if they were targeted by a genetic lesion (SNV, structural variation or focal CNA; ≤5 genes) in the infant MLL-R discovery cohort as determined by whole-genome sequencing (157/232 genes) or (ii) if they were annotated within a recurrently mutated pathway or if they had been described to be of importance for cancer (59 genes). This set of genes (n = 216) was analyzed by targeted capture. In addition, the coding exons from 16 genes were sequenced by Sanger sequencing (ABL1, AKT3, ALK, FOXO1, FOXO3, FOXO4, FOXO6, INPP5D, JAK2, MET, PDGFRα, PDGFRβ, PTEN, RAF1, RET and SYK), and putative SNVs and indel variants were detected by SNPdetector28. Non-silent sequence mutations were selected for validation by Sanger sequencing of both the tumor and matching normal samples (where available).

Experimental validation of somatic mutations and structural variants. Validation of the identified somatic tier 1 SNVs was performed using PCR amplification of the leukemia and germline DNA samples followed by Sanger sequencing (SJINF010 and SJINF019) or 454-based (SJINF006-009, SJINF011-018 and SJINF020-022) sequencing. We validated the tier 1–4 mutations by array-based capture followed by Illumina-based sequencing for INF001-005. Structural variations were validated by PCR and subsequent Sanger sequencing on whole genome–amplified DNA or by capture-based methods. The 454-based sequencing was performed as previously described21. Briefly, PCR products were subjected to library construction followed by 454 Titanium sequencing. Read sequences and quality scores were extracted with sffinfo (454 proprietary software), and sequences were aligned to NCBI Build 36 or 37 using SSAHA2 with the SAM output option. Alignments were imported to bam format using SAMTools. A SAMTools pileup file was generated, and read counts were determined by VarScan26. In the analysis of the 454 reads, a minimum base quality of 15 was required, with at least 20 reads aligned, to report the allele frequencies.

Oligonucleotide primers for genomic PCR were designed using the flanking sequences of each structural variation or SNV in Primer 3 (ref. 60). For SJINF001–SJINF005, SJINF0011R and SJINF002R, the structural variations were also validated using targeted capture followed by Illumina sequencing. Of the structural variations reported here, 77% have been experimentally validated.

Experimental validation of predicted in-frame fusion transcripts. Predicted in-frame fusion transcripts were validated by RT-PCR followed by Sanger sequencing of the purified PCR products. Primer sequences are available upon request.

Clonal evolution analysis. To decipher the potential clonal evolutionary path between the diagnostic and relapse tumor, we compared the MAFs of somatic SNVs detected at diagnosis and relapse. To have an accurate MAF readout, we performed deep sequencing using custom capture followed by Illumina sequencing on all tier 1–3 mutations detected by whole-genome sequencing in the trio of diagnostic, relapse and germline samples for cases SJINF001 and SJINF002. The median coverage of targets in the primary tumors SJINF001_D and SJINF002_D and in the relapse tumors SJINF001_R and SJINF002_R was 1,534×, 894×, 769× and 882×, respectively. The following criteria were used when assessing SNVs in the diagnostic tumor, requiring (i) that the mutant allele be missing from the matched germline sample and (ii) that the MAF be significantly different between the germline and diagnostic samples (Fisher’s exact test P ≤ 0.05). SNVs located on the X or Y chromosomes or on chromosomal segments with CNAs were removed. The remaining tier 3 SNVs were used to perform normal mixture modeling using the multclust package (version 3.4.10) in R. The optimal model is determined by Bayesian information criterion (BIC).

FLT3 and PIK3R1-PIK3CA modeling. The structure for autoinhibited, inactive FLT3 kinase domain was used to model the Asn676Lys substitution (Protein Data Bank (PDB), 1RJB). The structures for phosphorylated (PDB, 2PSO) and unphosphorylated (PDB, 2VPF) FGFR2 were used to model the mechanism of kinase activation61,62. The Asn676Lys substitution was introduced into the FLT3 kinase domain structure and energy minimized with default parameters in FOLDX53. Images and structure alignments were performed in PyMOL. The structure for PIK3R1-PIK3CA was used to interpret the insertion and deletion mutation sites (PDB, 3HIZ)64; this structure represents the inhibited state of the PIK3CA kinase domain. Prediction of disordered residues was performed by DISORED2 (ref. 65), and these residues were then visualized in PyMOL.

Clinical correlations. Association of the presence of an activating tyrosine kinase-pi3k-ras signaling pathway mutation or the presence of a KRAS or NRAS mutation with clinical variables was tested using the exact χ2 test. Event-free survival and overall survival distributions in different genetic groups were compared using the log-rank test. The cumulative incidence of any relapse was calculated according to the method of Kalbfleisch and Prentice and compared across genetic groups using Gray’s test66,67. We limited the outcome analyses to infants with MLL-R leukemias treated at St. Jude Children’s Research Hospital (n = 33 for 10-year overall survival and 10-year event-free survival; n = 31 for risk of relapse).

Pathway analysis. A genomic random interval (GRIN)68 model was used as a null model to evaluate the statistical significance of the pattern of overlap of genomic lesions with the loci of individual genes, predefined sets of genes and each base-pair locus in the genome. The null model represented each lesion as an interval of fixed length and random location that might occur at any location along the chromosome with equal probability.

Other statistical analysis. Unless otherwise specified, statistical tests were performed using R (version 3.1.1). A two-sided Fisher’s exact test was used to compare mutation frequencies in different patient groups. A two-sided Wilcoxon rank-sum test available in the R coin package (dealing with ties and reporting asymptotic P values) was used to compare the numbers of mutations, background mutation rates (BMRs) and CNAs in infant MLL-R ALL to those in non-infant MLL-R leukemias and to compare age in the two patient groups. SNVs were amplified from human cDNA, and sequence variants were amplified from human cDNA and, sequence variants (FLT3 D835Y (control), FLT3 D600H, FLT3 N676K, FLT3 D839G and FLT3 P934L; PIK3CA E542K (control) and PIK3CA H1047R (control); and PIK3R1

DNA constructs and retrovirus production. Full-length wild-type FLT3, PIK3CA and PIK3R1 were amplified from human cDNA, and sequence variants (FLT3 D835Y (control), FLT3 D600H, FLT3 N676K, FLT3 D839G and FLT3 P934L; PIK3CA E542K (control) and PIK3CA H1047R (control); and PIK3R1

doi:10.1038/ng.3230 NATURE GENETICS
Q572* (control)) were introduced using the QuikChange II XL Site-Directed Mutagenesis kit (Stratagene). Sequences encoding FLT3-ITD, PIK3R1-ITD (INF018), PIK3R1-R1α* (INF070) and PIK3R1-R1b (INF070) were cloned from primary patient material. All mutations were verified by sequencing the complete ORF of FLT3, PIK3CA or PIK3R1. Mutant cDNAs were cloned into defective mouse stem cell virus (MSCV) constructs coexpressing our mutants (wild-type FLT3, FLT3-ITD, FLT3 D835Y; FLT3 D600H, FLT3 N676K, FLT3 D839G or FLT3 P934L; wild-type PIK3R1, PIK3R1-ITD, PIK3R1-R1α*, PIK3R1-R1b or PIK3R1 Q572*) and GFP or an MSCV construct coexpressing the mutant (wild-type PIK3CA, PIK3CA E542K or PIK3CA H1047R) and mCherry. Full-length protein expression was verified by immunoblotting. Retroviral supernatants were produced using the ecotropic Phoenix packaging cell line (G.P. Nolan, Stanford University) and used to transduce BaF3 cells.

**BaF3 cell culture and cytokine-independent growth assays.** BaF3 cells were cultured in RPM1-1640 supplemented with 10% FCS and 10 ng/ml IL-3 (PeproTech). To analyze cytokine-independent growth, cells were transduced with the retroviral constructs, flow sorted with the mutant (wild-type PIK3CA, PIK3CA E542K or PIK3CA H1047R) and the membrane. α between PI3K

**Immunoblot analysis.** IL-3 (PeproTech) was withdrawn from BaF3 cells transduced with the FLT3 or PIK3 constructs overnight, and cells were serum starved for 6 h in medium with 0.3% FCS (HyClone). For FLT3 constructs, cells were resuspended in 1 ml of medium with or without 100 ng/ml FLT3 ligand (GenScript) for 5 min, washed once in ice-cold PBS and lysed with 1× RIPA (Cell Signaling Technology) containing Protease Inhibitor Cocktail (Sigma-Aldrich), PhosSTOP (Roche) and PMSF (Sigma-Aldrich). Immunoblotting was performed according to standard protocols. In brief, 50 µg of protein was separated on a 10% Bis-Tris gel (Life Technologies) and transferred to PVDF membrane (Life Technologies). We used antibodies to FLT3 (8F2), phosphorylated FLT3 (Y591; 54H1), STAT5 (9363), phosphorylated STAT5 (Y694; C11CS), ERK1/2 (137F5), phosphorylated ERK1/2 (T202/Y204; 20G11), GAPDH (6C5, Santa Cruz Biotechnology), PIK3CA (425S), PIK3R1 (4292), AKT (40D4), phosphorylated AKT (S473; D9E), pS6 (5G10) and phosphorylated S6 (S240/244; 2215). All antibodies, except for GAPDH, were purchased from Cell Signaling Technology. The antibodies to phosphorylated protein were used at a 1:500 dilution, the antibodies to total protein were used at a 1:2,000 dilution and control antibody (to GAPDH) was used at a 1:10,000 dilution.

50. Ding, L. et al. Genome remodeling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
51. Mardis, E.R. et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
52. Flicek, P. et al. Ensembl 2011. *Nucleic Acids Res.* **39**, D800–D806 (2011).
53. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank. *Nucleic Acids Res.* **37**, D26–D31 (2009).
54. Mardis, E.R. et al. SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS Comput. Biol.* **1**, e53 (2005).
55. Koboldt, D.C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
56. Rozen, S. & Skaltsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
57. Berman, H.M. The Protein Data Bank: a historical perspective. *Acta Crystallogr. A* **64**, 88–95 (2008).
58. Griffith, J. et al. The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. *Mol. Cell* **13**, 169–178 (2004).
59. Kalbfleisch, J.D. & Prentice, R.L. *The Statistical Analysis of Failure Time Data* (Wiley, 1980).
60. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. & Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Mol. Cell* **337**, 635–645 (2004).
61. Mandelker, D. *The FoldX web server: an online force field*. *Methods Mol. Biol.* **132**, 365–386 (2000).
62. Mandelker, D. et al. A frequent kinase domain mutation that changes the interaction between PI3Kα and the membrane. *Proc. Natl. Acad. Sci. USA* **106**, 16996–17001 (2009).
63. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. & Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
64. Gray, R.J. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Stat.* **16**, 1141–1154 (1988).
65. Kalbfleisch, J.D. & Prentice, R.L. *The Statistical Analysis of Failure Time Data* (Wiley, 1980).
66. Pounds, S. et al. A genomic random interval model for statistical analysis of genomic lesion data. *Bioinformatics* **29**, 2088–2095 (2013).