Developmental timing of mutations revealed by whole-genome sequencing of twins with acute lymphoblastic leukemia

Yussanne Ma[a,1], Sara E. Dobbins[a,1], Amy L. Sherborne[a,1], Daniel Chubb[a], Marta Galbiati[b], Giovanni Cazzaniga[b], Concetta Micalizzi[c], Rick Tearle[d], Amy L. Lloyd[a], Richard Hain[e], Mel Greaves[f,2,3], and Richard S. Houlston[a,2,3]

[a]Molecular and Population Genetics, Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey SM2 5NG, United Kingdom; [b]Centro Ricerca Tettamanti, Clinica Pediatrica, Università di Milano-Bicocca, Ospedale San Gerardo, 20900 Monza (Mi), Italy; [c]Experimental Clinical Hematology Unit, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) G. Gaslini, 16148 Genova, Italy; [d]Complete Genomics, Inc., Mountain View, CA 94043; [e]Paediatric Palliative Medicine, Children's Hospital for Wales, University Hospital of Wales, Cardiff CF14 4XW, United Kingdom; and [f]Haemato-Oncology Research Unit, Division of Molecular Pathology, Institute of Cancer Research, Surrey SM2 5NG, United Kingdom

Edited* by Max D. Cooper, Emory University, Atlanta, GA, and approved March 5, 2013 (received for review December 10, 2012)

Acute lymphoblastic leukemia (ALL) is the major pediatric cancer. At diagnosis, the developmental timing of mutations contributing critically to clonal diversification and selection can be buried in the leukemia's covert natural history. Concordance of ALL in monozygotic, monochorionic twins is a consequence of intraplacental spread of an initiated preleukemic clone. Studying monozygotic twins with ALL provides a unique means of uncovering the timeline of mutations contributing to clonal evolution, pre- and postnatally. We sequenced the whole genomes of leukemic cells from two twin pairs with ALL to comprehensively characterize acquired somatic mutations in ALL, elucidating the developmental timing of all genetic lesions. Shared, prenatal, coding-region single-nucleotide variants were limited to the putative initiating lesions. All other nonsynonymous single-nucleotide variants were distinct between tumors and, therefore, secondary and postnatal. These changes occurred in a background of noncoding mutational changes that were almost entirely discordant in twin pairs and likely passenger mutations acquired during leukemic cell proliferation.

fusion gene | copy number variants

The sharing in monozygotic (MZ) twins of identical but nonconstitutive and clone-specific fusion gene sequences (e.g., *ETV6-RUNX1*) in acute lymphoblastic leukemia (ALL) provided the first unambiguous evidence that genetic lesions, generated by chromosomal translocation, arise in utero (1). These data were interpreted to suggest that *ETV6-RUNX1* is likely to be a critical initiating lesion for *ETV6-RUNX1*–positive ALL, a view supported by single-cell genetic analysis (2) and modeling with murine (3) or human cells (4). However, such fusions are detectable in cord blood from newborn infants at rates ∼100-fold higher than the incidence of ALL, suggesting an obligatory requirement for additional mutations in leukemia development (5). Recurrent copy number variations (CNVs), mostly deletions, in ALL have been revealed by SNP arrays (6) and in twins with concordant ALL with *ETV6-RUNX1*. These CNVs are distinctive between twins of a pair, indicating a secondary, postnatal origin (7) that is also supported by single-cell clonal analysis (2). An additional layer of genetic complexity in acute leukemia is now apparent from genome sequencing (8, 9), and applying this technology to MZ twins with ALL provides a unique opportunity to decipher the timeline of all acquired genetic events in leukemogenesis.

Predicated on the hypothesis that shared identical mutations are prenatal in origin and twin-specific mutations are likely to be secondary and postnatal, we performed whole-genome sequencing of two MZ twin pairs with ALL.

Results

We studied two pairs of twins, twins 1.A and 1.B with *ETV6-RUNX1* fusion-positive ALL (7, 10) and twins 2.A and 2.B with *ETV6-RUNX1* fusion-negative ALL. Sequencing of matched tumor-normal (remission) samples from each patient was carried out using unchained combinatorial probe anchor ligation chemistry on arrays of self-assembling DNA nanoballs (11). Paired-end reads were aligned to the Human Genome [National Center for Biotechnology Information (NCBI) Build 37], resulting in an average coverage of 58.0–63.0 with 93–96% of bases called (Table S1). High-genotype calling accuracy in samples was confirmed by Illumina Omni1-Quad BeadChip (Illumina) SNP genotypes, with <0.03% discordance in each sample consistent with expected error rates for genotyping (Table S2). Comparison of germ-line variation between each identical twin pair provided confirmation of the estimated sequencing error rate of 1×10^{-5}.

In twin 1.A and 1.B, six identical interchromosomal rearrangements were identified (Fig. 1 and Table S3), but the only identified fusion product was *ETV6-RUNX1*. The 5q32 and 9p13.3 breakpoint featured a 111-bp insert mapping to chromosome 12, suggesting the complex *t* (9, 18, 12;21) translocation preceded the insertion of a section of 5q between 9p and 18p. The *t* (9, 18, 12;21) rearrangement did not have the typical configuration of the *ETV6/RUNX1* translocation generally documented; chromosome 21 sequence was identified on der(9) but not on der(12) as expected. Together with the presence of a section of 5p between 18q and 9p in the hybrid chromosome, this is compatible with the *ETV6/RUNX1* fusion being the initial translocation. It is likely that the complex rearrangement is a consequence of a cascade of chromosomal breakages and fusions initiated in a single cell in utero, presumably occurring in a short time frame. We assume that a ETV6-RUNX1 protein is the key functional product of this complex rearrangement but cannot rule out contributions from other components. This complexity of gene fusion is unusual in ALL but was suspected in this pair based on karyotype and FISH data (10) and is similar to chromosomal breakage/fusion cascades involved in generating the *TMPRSS2-ERG* fusion gene in prostate cancer (12).

We next examined evidence for CNV and loss of heterozygosity (LOH) in the twin pairs (Fig. 1, Table S4, and Fig. S1). In

Author contributions: M. Greaves designed research; Y.M., S.E.D., A.L.S., and A.L.L. performed research; M. Galbiati, G.C., C.M., R.T., and R.H. contributed new reagents/analytic tools; Y.M., S.E.D., A.L.S., and D.C. analyzed data; and M. Greaves and R.S.H. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]Y.M., S.E.D., and A.L.S. contributed equally to this work.

[2]M. Greaves and R.S.H. contributed equally to this work.

[3]To whom correspondence may be addressed. E-mail: mel.greaves@icr.ac.uk or richard.houlston@icr.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1221099110/-/DCSupplemental.

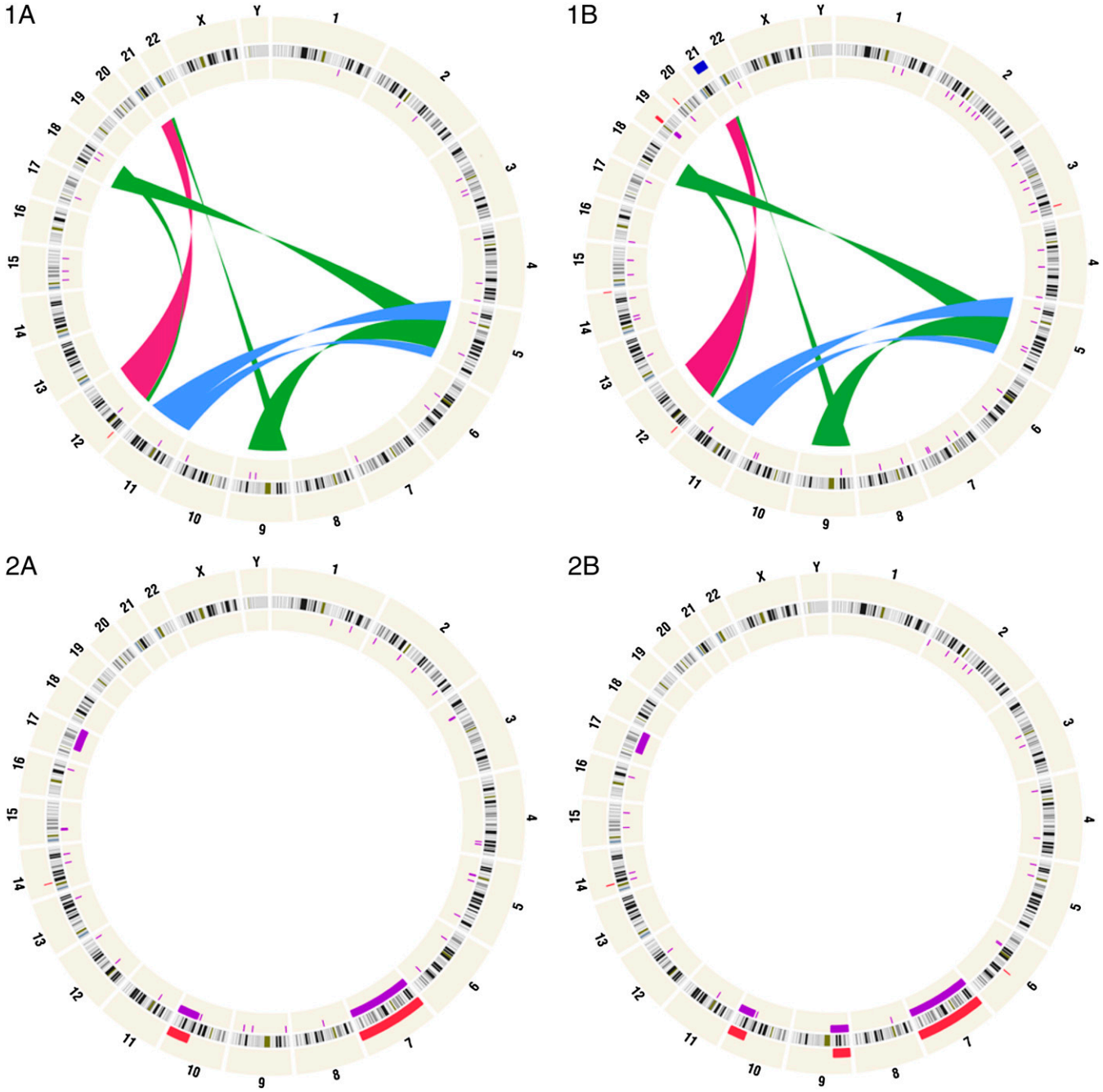

Fig. 1. Summary of lesions found in acute lymphoblastic leukemia. Shown are Circos plots (19) of somatic mutations in the each of the twin pairs. CNVs and LOH are depicted for each twin with regions of gain shown in blue, loss in red, and LOH in purple. Validated structural variations are depicted as colored ribbons in the interior of the plot. Panels refer to twin pair 1 and twin pair 2, and patients A and B in each twin pair.

twins 1.A and 1.B, deletions were observed at 1 and 7 cytobands, respectively, but none was shared, in accordance with previous array data (7). The majority of 21q was duplicated in twin 1.B only. Leukemic cells from both twin 2.A and 2.B had one copy of chromosome 7, a large portion of 10q and a region of 14q11.2 deleted. Two deletions in twin 2.B were distinctive (Table S4*A*).

We identified 27 and 41 cytoband regions with LOH events in twins 1.A and 1.B, only one common to both twins at 4q35.2. In twin pair 2, an inactivating germ-line neurofibromatosis type 1 (*NF1*) mutation 2763insA was identified, confirming the clinical diagnosis of neurofibromatosis (Table S5). In both twins, copy-neutral LOH at 17q encompassing *NF1* resulted in homozygosity for 2763insA (Fig. 1 and Table S4*B*). CNVs and LOH identified through sequencing confirmed results from the SNP array analysis (Fig. S2).

The mutation rates in the tumors from both sets of twins were lower than that seen in solid tumors (0.25–0.44 per megabase; Table S6). There was a paucity of single-base/small-indel driver mutations, consistent with the absence of strong positive selection. No high-confidence coding changes, including short insertions and deletions, were identical between either twin pair. However, we identified four missense SNVs in twin 1.A, in *ABLIM1*, *KLHL4*, *ZFHX3*, and *PAPPA* (Table S7). Twin 1.B also harbored a somatic mutation in the intron of *PAPPA*. In twin1.B, we identified a nonsense mutation in *ETV6* (R378X) and

missense mutations in *CAMTA1*, *CAPN14*, *CCDC136*, *DQX1*, *GOSR1*, *IKIP*, *METTL11B*, *PCDH11X*, *RAMP3*, *TRIM16*, and *IGSF3*; Table S7). In twin 2.A, we identified a deletion in *LOC283685*, frameshift mutations in *ANKRD31*, *LOC100128838*, and *LOC100130084*, and missense mutations in *OR4C3*, *OR51B5*, and *SLAMF7*. In twin 2.B, insertions were identified in *TTF1* and *TMEM132B* and missense mutations in *FAT3*, *LAMA1*, *MAP3K6*, *PIK3R3*, *SETD2*, *GRM1*, and *SYNM*. Strikingly, none of these coding region nonsynonymous (ns) mutations were shared within twin pairs.

Noncoding regions, inter- and intragenic, were also analyzed for point mutations in the twin pairs. Six identical noncoding point mutations in twin pair 1 and five in twin pair 2 were identified and confirmed by Sanger sequencing (Table S8); none of which had known encyclopedia of DNA elements (ENCODE) annotations (http://encodeproject.org). Four of the identical mutations in twin pair 1 were intronic to *RDH14*, *MID1*, *YPEL2*, and *KIAA1680*, all genes with no known role in leukemia, or cancer in general. Two shared mutations were intergenic on chromosomes 12 and 20 (Table S8). Four of the identical mutations in twin pair 2 were intronic to *FAT3*, *HIST1H4K*, *XDH*, and *C21orf67*, and one intergenic within chromosome 2 (Table S8). *In silico* analysis of these genomic regions revealed that the *KIAA1680* and *MID1* mutations map within regions encoding putative noncoding RNA. In twin 2.B, an additional intronic mutation was also identified in *FAT3*.

Performing a gene-centric analysis identified seven genes affected by somatic change in twin 1.A [one by CNV, two by LOH, and four by single-nucleotide variant (SNV) or small indel] and 427 in twin 1.B (400 CNV, 45 LOH, 12 SNV/small indel; Table S7). *ETV6* was affected by a somatic nonsense mutation in twin 1.B and copy number change in twin 1.A (Table S7). A total of 2,076 genes were altered by somatic change in twin 2.A (1944 CNV, 137 LOH, 7 SNV/small indel) and 2,610 in twin 2.B (2,154 CNV, 964 LOH, 9 SNV/small indel). A total of 1,948 genes were altered in both twin 2.A and twin 2.B.

Discussion

The biology of leukemia in twins is no different to that in singletons. Data derived from twin studies can therefore inform an understanding of ALL in its normal clinical setting. Although the depth of sequence coverage in our study was high, we acknowledge that clonal dominance of the presenting leukemia could potentially mask sequence changes and lead to a failure to demonstrate subclonal mutations between twin pairs. However, if any somatic mutation preceded the *ETV6-RUNX1* fusion, then as with the fusion gene itself, we would expect it to be retained in all clonal descendants. Accepting the caveat, our data (summarized in Fig. 2) indicate a relatively simple developmental sequence for coding changes in twin pair 1 with an identical, albeit complex, *ETV6-RUNX1* fusion event occurring in utero. A total of 16 nsSNVs and 8 CNVs were present in twin pair 1, but none of these 22 mutations was shared, indicating that these were secondary genetic changes and almost certainly postnatal in origin. In this pair, therefore, *ETV6-RUNX1* is the only early (prenatal) protein encoding mutation, endorsing its likely role as an initiating lesion.

Similarly, in twin pair 2, a total of 16 nonshared nsSNVs, presumably postnatal in origin, were present. However, in addition to the constitutive *NF1* mutation, three other indistinguishable

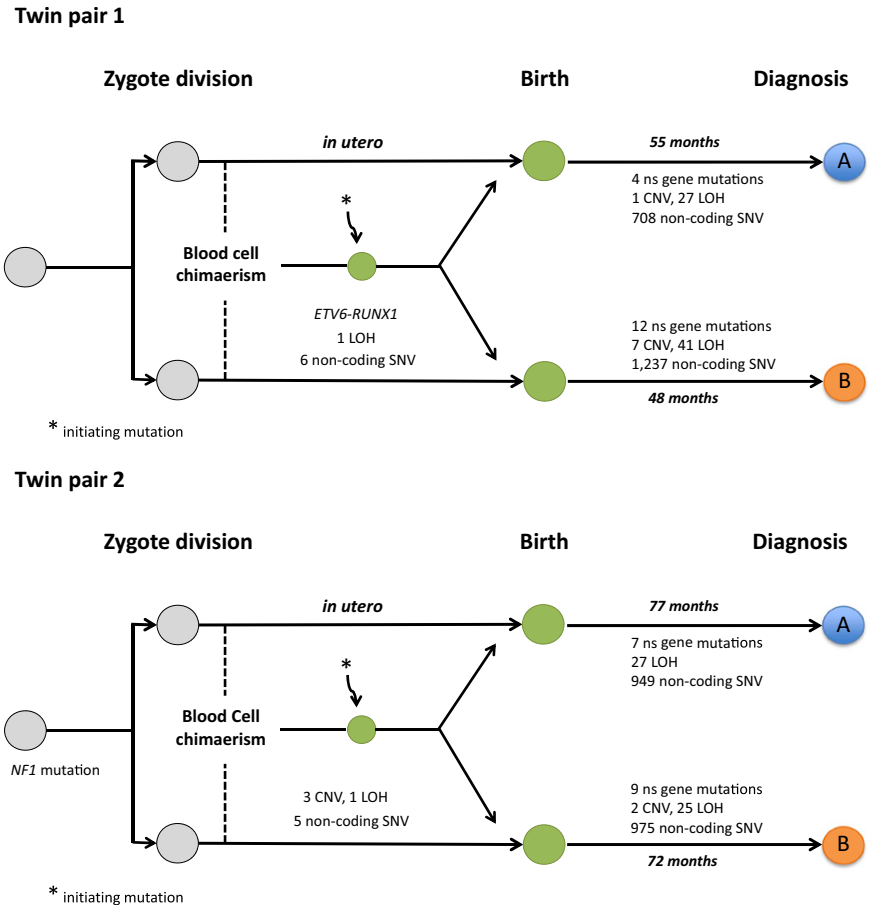

Fig. 2. Developmental timeline of acute lymphoblastic leukemia in the two twin pairs revealed by whole-genome sequencing.

somatic chromosome changes, or CNVs (deletion of chromosomes 7, 10q, and 14q11.2), were present in the leukemic cells of both twins along with copy-neutral LOH of 17q. We conclude, therefore, that these events most probably occurred in utero in the same single clone of cells, but given the lack of uniqueness of these genetic changes, we cannot exclude that they arose independently and postnatally.

The very small number of genome-wide noncoding somatically acquired nonconstitutional SNVs that were shared (six in twin pair 1, five in twin pair 2) must have occurred in utero, in a single clone of cells, subsequently shared by intraplacental dissemination (1). Shared SNVs in twin pair 2 confirms a common clonal origin for their ALLs. The functional relevance, if any, of these mutations in leukemic cells is, however, uncertain in the absence of data on recurrence in a larger series of cases. None of them has ENCODE annotations (13), which might have highlighted potential gene regulatory functions. Random mutations of no functional relevance to leukemogenesis accumulate in normal hematopoietic stem cells and progenitors increasing in frequency with age and proliferative history (14). The small number of genome-wide sequence changes in common between twin sets may then most likely reflect the accrual rate of passenger mutations in the prior life history of the fetal cell transformed by either *ETV6-RUNX1* (twin pair 1), *NF1* inactivation, and other genetic events (twin pair 2). The substantially greater number (~1,000 per case) of noncoding mutations, unique to each case, we presume are most likely "passengers" accumulated during the proliferative expansion of the leukemic clones postnatally.

Only a fraction of the coding and nsSNVs in both twin pairs involve genes recurrently mutated in leukemia or cancer (Table S7). Although these mutations could still have functional relevance, these data suggest that, in common with other pediatric cancers (15–17) and adult acute myeloid leukemia (AML) (14), childhood ALL can develop in the absence of genetic instability and with only a small number of driver mutations. Indeed, the relatively high recurrence of CNV in ALL (6) suggests that they, rather than SNVs, are more highly selected as driver lesions to complement the *ETV6-RUNX1* initiating lesion.

Materials and Methods

Patient Samples. Two pairs of monozygotic twins with pediatric B-cell precursor ALL were studied. Twins 1.A and 1.B had been diagnosed with *ETV6-RUNX1* fusion-positive ALL at ages 55 and 48 mo, respectively. Twins 2.A and 2.B, who had a clinical diagnosis of NF1, had been diagnosed with *ETV6-RUNX1* fusion-negative ALL at 77 and 72 mo, respectively. DNA was extracted from leukemic bone marrow acquired at diagnosis and peripheral venous blood during remission from each patient using standard methods and quantified using PicoGreen (Invitrogen). Remission bloods were assessed for minimal residual disease (MRD) by PCR with immunoglobulin heavy chain and T cell receptor consensus primers, and each twin had a normal blood profile with an MRD level of 10^{-4} (i.e., <1 leukemic cell per 1,000), thereby excluding contamination likely to impact on sequencing. Ethical review committee approval for the study was obtained from the Royal Marsden NHS Hospitals Trust; University Hospital of Wales, Cardiff; and Ospedale San Gerardo, Italy.

Genome-Wide SNP Genotyping. Genome-wide genotyping was conducted using an Illumina Omni1-Quad BeadChips according to the manufacturer's protocols (Illumina; www.illumina.com). Genotypes were called using Illumina BeadStudio software.

Sanger Sequencing. All translation breakpoints and putative SNVs shared between twin pairs identified by whole-genome sequencing were subjected to Sanger confirmation in samples from both individuals. PCR products were directly sequenced using dye-terminator chemistry implemented on ABI-3370xl semiautomated sequencers (Applied Biosystems). Sequences were visualized using Mutation Surveyor Software (SoftGenetics). Details of the PCR and sequencing primers are available on request.

Whole-Genome Sequencing and Somatic Variant Detection. Sequencing of matched normal and tumor samples from each twin pair was carried out using unchained combinatorial probe anchor ligation chemistry on arrays of self-assembling DNA nanoballs (11). The gross mapping yields were 186–196 and 180–192 Gb for the four normal and four tumor samples, resulting in an average coverage of 59.6–63.0 and 58.0–61.3 haploids (Table S1). To identify sequence variation in each sample paired-end reads were aligned to Human Genome NCBI Build 37. Variations between the reference genome and each of the samples were called and scored using a local de novo assembly algorithm (11). In addition, a reference score was calculated for each called base in the genomes (11).

To eliminate somatic SNVs most likely to be false positives attributable to systematic error in the sequencing and variant calling, we applied a number of quality control filters on the SNVs identified by Complete Genomics (CG) (www.completegenomics.com). Somatic SNVs were only classified as "high confidence" if they were not present in more than one healthy Centre d'Etude du Polymorphisme Humain sample (public data provided by CG), if there was not more than one other variant within 100 bp, if they occurred in regions with a mappability score greater than 0.5 [defined by University of California, Santa Cruz (UCSC) mappability tracks; www.genome.ucsc.edu] (i.e., reads mapped to this region could only be mapped to one other region in the genome), if they were in a region with coverage greater than 20 and if they were supported by at least four uniquely mapped good-quality reads (Fig. S2). A similar procedure was adopted for identifying germ-line variants. Sequence variants were annotated using tools and data provided by UCSC, Annovar (www.openbioinformatics.org/annovar), Human Genetics Mutation Database (www.hgmd.cf.ac.uk), Biobase (www.biobase-international.com/product/genome-trax), GeneCards (www.genecards.org), and CG.

Copy Number Variation. Using next generation sequencing data, we identified regions of CNV in all four tumor DNAs relative to the matched normal using algorithms developed by CG. Genome-wide coverage was smoothed in 100-kb windows, corrected for GC content, and normalized using composite baseline coverage from multiple healthy samples. CNV levels were called using a Hidden Markov Model (HMM). A CNV event was defined as a block with coverage level of less than 2 (deletion) or greater than 2 (duplication). Somatic CNVs are reported by cytoband and position, disregarding CNV events with reported CNV events in the normal genome of each sample which are ascribed to bad alignment. CNVs close to centromeres or at the beginning or end of chromosomes were also discarded.

Regions of CNV were confirmed using Illumina Omni1-Quad BeadChip array data by implementing PennCNV software (www.openbioinformatics.org/penncnv) (18). Raw intensity values were processed to obtain normalized allele-specific values for each probe using Illumina Bead Studio Software (Illumina). B-allele frequency (BAF) and log R ratio (LRR) values were calculated within PennCNV package implementing a correction for GC-content bias. LRR values for somatic CNVs were calculated from the difference of the tumor and normal LRR values and segmented copy number calls of 0, 1, 2, 3, or 4 derived using an HMM. Only CNVs with supporting evidence from array data are reported.

LOH. An LOH event was defined in the SNP array data as an SNP that was heterozygous in the normal sample and homozygous in the tumor sample. BAF in the tumor sample was examined in genomic regions containing SNPs with LOH to estimate the boundaries of the LOH.

To identify LOH events from CG data, we scanned the genome in 100-kb windows for LOH sites. A region was defined as having an LOH event if it contained 10 or more LOH sites and had at least five times as many LOH sites than heterozygous sites in the tumor sample and at least five times as many LOH sites than potentially erroneous sites. The strength of evidence for LOH was reduced for twin 1.A and 2.A, potentially indicating the presence of clonality (and, therefore, residual heterozygosity). For these twins, we called large regions of LOH allowing for higher levels of heterozygosity in the tumor sample. LOH was also called using a gene-centric approach requiring a minimum of five LOH events for each gene and at least five times the number of LOH events as remaining heterozygotes. A gene is only classified as affected by LOH where a damaging nonreference variant was also present in the gene.

Large-Scale Structural Variation. We used the mate-paired data to identify putative large-scale structural variations (SVs). De novo assembly around putative SV breakpoints was conducted using CG algorithms and SVs present in the patient's germ-line genome were assumed to be alignment error and were excluded. We applied a number of criteria to eliminate false positives. SV events were taken forward where there were more than 10 mate pairs in a cluster, de novo assembly of the junction was successful, there was a high

mapping diversity, and specific repeat sequences on the left and right side of the junction were absent.

Conventional cytogenetics of leukemic clones was conducted using standard karyotyping methodologies, and standard criteria to define a clone were applied.

ACKNOWLEDGMENTS. This work was supported by Leukaemia and Lymphoma Research Grants LLR100201 and LLR11021, Kay Kendall Leukaemia Fund Grant KKL 531, and Cancer Research UK Grant C1298/A8362 (supported by the Bobby Moore Fund).

1. Greaves MF, Maia AT, Wiemels JL, Ford AM (2003) Leukemia in twins: Lessons in natural history. *Blood* 102(7):2321–2333.
2. Anderson K, et al. (2011) Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* 469(7330):356–361.
3. van der Weyden L, et al. (2011) Modeling the evolution of ETV6-RUNX1-induced B-cell precursor acute lymphoblastic leukemia in mice. *Blood* 118(4):1041–1051.
4. Hong D, et al. (2008) Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science* 319(5861):336–339.
5. Mori H, et al. (2002) Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci USA* 99(12):8242–8247.
6. Mullighan CG, et al. (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* 446(7137):758–764.
7. Bateman CM, et al. (2010) Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* 115(17):3553–3558.
8. Ding L, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481(7382):506–510.
9. Zhang J, et al. (2012) The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* 481(7380):157–163.
10. Broadfield ZJ, et al. (2004) Complex chromosomal abnormalities in utero, 5 years before leukaemia. *Br J Haematol* 126(3):307–312.
11. Drmanac R, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327(5961):78–81.
12. Berger MF, et al. (2011) The genomic complexity of primary human prostate cancer. *Nature* 470(7333):214–220.
13. Dunham I, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
14. Welch JS, et al. (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell* 150(2):264–278.
15. Parsons DW, et al. (2011) The genetic landscape of the childhood cancer medulloblastoma. *Science* 331(6016):435–439.
16. Zhang J, et al. (2012) A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature* 481(7381):329–334.
17. Molenaar JJ, et al. (2012) Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* 483(7391):589–593.
18. Wang K, et al. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17(11):1665–1674.
19. Krzywinski M, et al. (2009) Circos: An information aesthetic for comparative genomics. *Genome Res* 19(9):1639–1645.