doi:10.1038/nature10725

METHODS

Illumina Library Construction/Illumina Sequencing

All methods in the library construction and whole genome DNA sequencing have been described previously[44,45]. Tumor and normal genomes were sequenced to at least 30-fold haploid coverage, with corresponding diploid coverage of at least 99.5%. Detailed information regarding runs and lanes generated for the twelve tumour/normal pairs is included in Supplementary Table 7.

Transcriptome sequencing

During library preparation, the RNA was purified by phenol/chloroform/isoamyl alcohol (Life Technologies) extraction followed by sodium acetate / ethanol precipitation after each enzymatic step. For library construction 2-5 μg of total RNA was DNAse I (Life Technologies) treated for 15 minutes at 23ºC, after which polyadenylated (poly-A) RNA was isolated with uMACS columns (Miltenyi Biotec). The isolated poly-A RNA served as a template for cDNA synthesis with random hexamers using the Superscript Double-Stranded cDNA Synthesis kit (Life Technologies). The resulting cDNA was fragmented by a Covaris model E210 according to the manufacturer's recommended conditions to generate fragments with a peak distribution of approximately 200bp. PAGE size selection was performed to further restrict the DNA size range to 100-300bp prior to DNA end repair, and the addition of library adapters with the NEB Next DNA sample prep kit (NEB). After 10 rounds of PCR amplification with primers PE 1.0 and PE 2.0 (Illumina), the final product was size selected by PAGE (290-325 bp). The resulting libraries were quantitated with the QPCR NGS library quantification kit (Agilent Technologies) using a PhiX control library (Illumina) as an external standard. Bridge PCR clusters were generated on a V4 PE flow cell (Illumina) using a CBOT (Illumina).

Flow cells were loaded onto an Illumina GAIIx for a paired end 2x101 cycle sequencing run using SCS version 2.8 software and SBS version 5 reagents, with the resulting base call files (bcl) converted to fastq format and used in the analysis pipeline.

Alignment, coverage, and quality assessment

All Illumina paired-end reads from both tumour and matched normal ends were aligned to the human NCBI Build 36 reference sequence using BWA (0.5.5) aligner[46,47], all lanes were merged and deduplicated using Picard 1.29 (http://picard.sourceforge.net). The genotype files from Affymetrix SNP 6.0 and 500K microarrays were used for coverage analysis and for QC purpose. A copy number profile of each lane was constructed and compared to those derived by SNP array to ensure consistency.

Effective coverage of the whole genome and whole exome was obtained by summarizing coverage of aligned bases with quality score >=15 at each position of the reference genome (excluding sequencing gaps and ambiguous bases) using the Coverage module of Bambino[48]. We defined a base "covered" if the "effective coverage" was at least 10x. We also excluded all ambiguous bases and sequencing gaps in hg18 from our coverage analysis. The exome annotation is based on NCBI RefSeq[49].

Detection of substitution (SNV) and insertion/deletion (indel) sequence mutations

SNVs and indels were analysed independently by Washington University Genome Institute (WU) and St. Jude Children's Research Hospital (SJCRH) using different approaches. The results generated from the two institutes were combined and sent for validation to generate the final candidate SNV and indel list for experimental validation.

At WU. all reads were aligned using BWA (version 0.5.5 by lane), lanes were merged and deduplicated using Picard 1.07 (http://picard.sourceforge.net) and then variants called using Samtools (svn rev 454)[46]. Somatic single nucleotide variants were detected as previously described[44], but using a program that directly accesses a BAM file called SomaticSniper (Larson *et al*. manuscript in preparation). High quality somatic predictions were those sites with a somatic score greater than 40 and an average mapping quality greater than 40. This represents a slight modification of the previous categorisation described by WU, which required a minimum average mapping quality of 70, as reads were aligned using BWA, and BWA calculates its mapping qualities differently than MAQ which was used in prior studies. The predicted SNVs are compared to the most current version of dbSNP[50] (build 129-130). For SNVs, we required both a positional and allele match. In addition we also compared the predicted SNVs to SNPs found in CEU and YRI trios as described in Ding *et al*.[44]. All predicted SNVs were filtered through a SNV false-positive filter developed at the Genome Institute that is based on a set of criteria including mapping quality score, average supporting read length, average position of the variant in the read, strand bias and the presence of homopolymers.

At WU, indels were called using a modified version of Samtools to identify indel mutations that were more likely to have occurred in the tumour than the normal. The following scheme was implemented to identify putative somatic indels: analogous to the approach used for SNVs, a comparison of the Samtools likelihoods of the indel in tumour and normal reads was performed to generate a somatic score representing the Phred scaled probability that each indel is somatic. Indels were removed where any of the following conditions was true: the somatic score was 0, the Samtools tumour and normal consensus calls were the same, the Samtools call in normal was not wild type, or the number of reads was greater than 100. Furthermore, for predictions of 1-2 base pairs in size, we performed a one-sided Fisher's exact test on the read counts supporting the indel in tumour and normal to test if the indel occurred at a lower frequency in the normal. One to two base pair indels where the *P* value was greater than 0.01 were removed. Finally, predictions found to be contained or adjacent to runs of base quality 2 bases were removed as these are set to indicate a failure in base calling during the Illumina Base Calling Pipeline and many indel predictions were found to arise solely from these reads. In addition, Pindel and GATK were used to detect indels[51,52]. A modified version of Pindel that reads directly from the BAM file was used. Insert size estimates required as input were derived from BWA. Tumour and normal reads were tagged and pooled, and somatic filtering was done by removing all calls which had even one normal read aligning to the same putative event as tumour reads. This modified version has since been incorporated into the original Pindel source code. GATK was run with all default parameters. This includes GATK IndelGenotype V2.0 with window size of 300 base-pairs.

Indels and SNVs were grouped into tiers based on genome annotation as described previously[44,45].

At SJCRH, putative sequence variants including SNVs and indels were initially detected by running the variation detection module of Bambino[48] using the following three parameters: (1) a high quality threshold for pooled tumour and matching normal bam files (min-quality 20 -min-flanking-quality 20 -min-alt-allele-count 3 -min-minor-frequency 0 -broad-min-quality 10 -mmf-max-hq-mismatches 4 -mmf-min-quality 15 -mmf-max-any-mismatches 6); (2) a low quality threshold for pooled tumour and matching normal bam files (-min-quality 10 -min-flanking-quality 10 -min-alt-allele-count 2 -min-minor-frequency 0 -broad-min-quality 10); and (3) a high tolerance for the number of mismatches for normal bam file alone (min-quality 20 -min-flanking-quality 15 -min-alt-allele-count 2 -min-minor-frequency 0 -mmf-max-hq-mismatches 15 -mmf-min-quality 15 -mmf-max-any-mismatches 20). In addition to Bambino, putative indels were also found by a *de novo* assembly process which constructs contigs using unmapped reads and then re-maps them to the reference genome followed by a Smith-Waterman alignment to detect indels. In this process, unmapped reads include (1) unmapped reads whose mate is mapped to the genome; (2) reads with indels in the CIGAR (C̲ompact I̲diosyncratic G̲apped A̲lignment

Report) string; (3) reads with at least 4 high-quality (quality value >=20) mismatches; and (4) reads with high-quality (quality value at least 20) soft-clipped bases in the CIGAR string. All putative sequence variants were further assessed to determine their accuracy and somatic origin using the processes described below. Velvet[53], BLAT[54] and SIM[55] were the three programs used for assembly, mapping, and Smith-Waterman alignment, respectively.

A putative somatic sequence mutation determined by a SJCRH process was collected based on the following criteria: (1) the variant site is absent in the normal-only analysis; (2) Fisher's exact test *P* value indicates that the number of reads harbouring the non-reference allele is significantly higher in tumour; (3) the non-reference allele frequency in normal is <=5%; and (4) mutant alleles are present in both orientations. Higher *P* value and absence of non-reference allele in normal is required for a variant to be considered somatic if it matches dbSNP build 130 or is located in an unmappable region (determined by recurrence of 75mers across the reference genome) or is inside a polynucleotide repeat. Substitution variants are classified into four categories based on combination of their *P* value and sequence quality scores: High quality, high *P* value; high quality, low *P* value; low quality, high *P* value; low quality, low *P* value. *P* value refers to the *P* value of Fisher's exact test comparing the distribution of the alternative allele in tumour and normal. High *P* value, *P*<0.05; low *P* value, 0.05<*P*<0.10. A final review process re-maps and re-aligns the reads harbouring the non-reference allele to the reference genome to filter potential false positive calls introduced by mapping in repetitive regions and alignment artefacts. For putative somatic indels, the review process re-aligns all reads in tumour and normal at the indel site to a mutant allele template sequence constructed by substituting the wild-type allele with the indel. Presence of reads in normal that cover the mutant allele is considered a germline variant.

Tier annotation for sequence variations

Transcripts from Ensembl[56] build (54_36) and Genbank[57] (build download May 21, 2009) were used for annotation. Variants were classified into the following four tiers. (1) Tier 1: Coding synonymous, nonsynonymous, splice site, and non-coding RNA variants; (2) Tier 2: Conserved variants (cutoff: conservation score 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 annotation (Regulatory annotations included are targetScanS, ORegAnno, tfbsConsSites, vistaEnhancers, eponine, firstEF, L1 TAF1 Valid, Poly(A), switchDbTss, encodeUViennaRnaz, laminB1, cpgIslandExt); (3) Tier 3: Variants in non-repeat masked regions; and (4) Tier 4: the remaining SNVs.

Validation of tier 1 somatic mutations

All tier 1 predicted somatic SNVs and indels including those found in poorly covered regions by whole-genome sequencing were validated by using either genomic PCR and sequencing using the Roche 454 (for all samples except SJTALL001) or Sanger sequencing (SJTALL001). A fraction of predicted mutations at the 5' or 3' UTR were also validated. Primer design and PCR amplifications were carried out as previously described[45]. The 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 then aligned to NCBI Build 36 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 VarScan. In the analysis of 454 reads generated by validation experiment, we required a minimum base quality of 15, with at least 20 reads aligned, to report the allele frequencies. Of the 228 validated somatic mutations, 50 (22%) had poor coverage in the initial whole-genome sequencing: i.e. the coverage of the tumour is lower than 16x or that of the matching normal is lower than 8x. The lowest covered sites include a NOTCH1 (R1598P) mutation with 5x coverage of tumour and 4x coverage of matching normal.

Analysis of background mutation rate in T-ALL cases

We followed the current convention which incorrectly assumes that the human genome is a 3-billion base haploid genome[58,59] instead of a ~6 billion-base diploid genome to calculate the background mutation rate for comparability with other studies. The analysis used validated silent somatic mutations as the non-functional background mutations. We obtained the total number of effectively covered coding bases (i.e. covered by >10x in both tumour and matching normal) in all RefSeq protein coding exons for each case. The background mutation rate is the silent mutation rate in coding region (the number of validated silent somatic mutations divided by the total effectively covered coding bases) adjusted by silent-to-non-silent ratio (estimated to be 0.350 by the TCGA Consortium) across the coding regions. Since our validation included mutations in poorly covered regions, only the validated silent mutations that have >10x coverage are included in this analysis.

Recurrence screening for sequence variations

We performed recurrence screening of a cohort of 94 T-ALL samples from SJCRH, the Children's Oncology Group and L'Associazione Italiana Ematologia ed Oncologia Pediatrica (AIEOP) (52 ETP and 42 non-ETP) for the following 42 genes: *BCL11B, BRAF, CBL, CTCF, DCLRE1C, DNM2, ECT2L, EED*, *EP300, ETV6, EZH2, FBXW7, FLT3, GATA1, GATA2,* *GATA3, HISTH1B, HNRPA1, HNRPR, IFNAR1, IFNAR2, IGF1R, IKZF1, IL7R, KRAS, JAK1, JAK2, JAK3, KDM5, NF1, NOTCH1, NRAS, PHF6, PTEN, PTPN11, RELN, RUNX1, SETD2, SH2B3, SMARCA4, SUZ12* and *TYK2.* The screening was done using PCR-based 3730 sequencing by WU and by Beckman Coulter Genomics as previously described[44,60]. PCR primer sequences are available upon request. Putative SNVs and indel variants were detected by SNPdetector[61] and PolyScan[62]. Novel, non-silent sequence mutations were selected for validation by sequencing both the tumour and the matching normal samples.

Detection of inherited sequence variations

Novel, non-silent coding germline variants were identified by the following process using the variation detection output generated by Bambino with a high quality threshold for pooled tumour and matching normal BAM files. (1) All variants with a non-reference allele covered in both forward and reverse orientation and with a non-reference allele fraction (NAF) exceeding 0.30 for high-quality normal reads or NAF exceeding 0.15 for all normal reads were retained as high-confidence candidate germline variants. (2) High-confidence germline variants that were not found in dbSNP were retained as novel variants. In addition, variants in dbSNP that were also present in OMIM or COSMIC[63] were retained as these variants are likely to be of biologic importance. (3) Novel variants that were found in coding regions of RefSeq were annotated and non-silent variants (missense, splice and nonsense) were retained. (4) Variants located in polymers (n>=8) or microsatellite repeats (repeat size ranging 2-5 and repeat unit >=6) were removed due to high rate of sequencing error. (5) Variants that show a drop in quality score compared with the flanking bases were removed. (6) Non-specific variants were identified by performing a BLAT search using the reads harbouring the non-reference against the reference human genome. Variants with no unique read support were considered to be caused by non-specific mapping and removed. (7) All reads were realigned to +/-100bp of the variant sites by SIM, which implements the Smith-Waterman algorithm. Variant detection was rerun using the alignments generated by SIM. This removed false positive variants caused by BWA mis-alignment. (8) Variants that were located in a SNP cluster (e.g. >=3 novel variants) were removed as they are likely to be unspecified repeats in human based on our evaluation of false variants found in male X chromosomes. (9) Variants detected in segmental duplication and that had abnormal coverage compared with the genome-wide average (FDR q-val <= 0.05) were removed.

Identification of structural variations using NGS data

Structural variations including inter-chromosomal translocations (CTX), intra-chromosomal translocations (ITX), inversions (INV), deletions (DEL), and insertions (INS) were analysed by CREST (Clipping REveals STructure), a novel algorithm that uses the soft-clipped reads to directly map the breakpoints of structural variations[5]. All samples were analysed using the paired analysis module, which filters SVs present in the matching normal sample. Two additional methods, BreakDancer[64] and Geometric Analysis of Structural Variants (GASV)[65], that use discordant paired-end reads to map structural variations were also run with modifications (described below) for comparison purposes.

BreakDancer was run using the default parameters. For each predicted SV, we first checked whether discordant mapping of paired-end reads was caused by repetitive regions in the human genome. All supporting reads were extracted in fastq format and each read re-mapped to the reference genome using BLAT. If a read-pair was mapped within the library insert range (mean insert size +/- 3 standard deviation), it was not considered to be a supporting read pair for the SV. All SVs with ≥3 supporting read pairs and a BreakDancer score ≥30 after the re-mapping were retained and the tumour-only SVs were considered to be putative somatic SVs. The putative somatic SVs were then subjected to an assembly process to evaluate their validity. All reads mapped within 1kb of the two breakpoints along with their unmapped mate pairs are extracted using the mapping information based on the bam files. We then ran phrap[66] to assemble the extracted sequences into contigs using base call, quality value and paired-end sequence information. Assembly was carried out in two iterations because the first iteration usually generated contigs that represent the wild-type allele unless the alternative allele was a homozygous genomic change. The second iteration began with reads not assembled in the first iteration, which generated contigs for the heterozygous alternative allele. All contigs were mapped to the reference human genome using BLAT. If a contig had two distinct parts (i.e. two regions with minimum overlapping) mapped to two different genomic regions with high similarity (≥97%) and good read-length (≥30bp), it was considered a cross-junction contig. Once such a contig is identified and there is no germline reads mapped to the breakpoints identified in the BLAT alignment, the SV is considered an assembly-validated somatic SV.

GASV (version 1.4) was run using the default parameters. Paired tumour/normal bam files were used to identify putative somatic SVs.

Experimental validation of structural variations

All structural variations were validated by Sanger sequencing. Oligonucleotide primers for genomic PCR were designed for the 1000bp flanking sequences of each SV using Primer 3 (ref. [67]). In two cases, a second iteration of primer design was carried out because there were multiple SVs detected within 1kb to account for the presence of a second SV in the flanking region.

Annotation of structural variations

SVs with at least one breakpoint in a gene coding region were further analysed for their validity to encode a fusion protein. Each predicted fusion transcript was defined as a list of exons. There were "normal" exons which correspond exactly to existing annotated exons, and there were fused exons which are produced by structural variation events with both breakpoints in exons. The sequence of the fused exons is determined using the assembly of reads that cross the breakpoints and the annotation of the exons. For each exon in the list, we calculated exon length, using the annotation for normal exons and sequence length for fused exons. Furthermore, we calculated the number of bases that each exon contributes to the CDS based on the annotated CDS start and end positions. The number of "CDS bases" was 0 for exons lying outside of the CDS start and stop, the full exon length for exons wholly contained between start and stop, and a portion of the exon length for those containing CDS start or stop. If the sum of the number of CDS bases is a multiple of 3, then the CDS was considered to be in-frame. If not, then it was considered out-of-frame.

Experimental validation of predicted in-frame fusion transcripts

Fusion transcripts were validated by RT-PCR and direct Sanger sequencing of purified PCR products as previously described[23]. Primer sequences for RT-PCR are shown in Supplementary Table 5.

Identification of copy number variations (CNVs) in whole genome sequencing data

CNVs were analysed independently by WU and SJCRH using different approaches. All CNVs that match T-cell receptors at the following loci were filtered from further analysis: 2p11.2 (*IGK@*), 7p14.1 (*TRG@*), 7q34 (*TRB@*), 14q11.2 (*TRA@*) and14q32.33 (*IGH@*). The final results combined the output generated from the two institutes.

At WU, an in-house developed tool, cnvHMM (unpublished), was used to detect copy number alterations genome wide for individual samples. The methods and rational behind this tool is described in Ding *et al*.[44]. The only difference from what was previously described is the window size used (1000 base-pairs) and the log likelihood ratio (LLR) cutoff used (LLR>= 100). For groups of samples, we used CMDS to identify recurrent DNA copy number changes[68].

At SJCRH, CNVs were identified by evaluating the number of sequence reads aligned at each base using the novel algorithm CONSERTING (COpy Number SEgmentation by Regression Tree In Next-Gen sequencing, manuscript in preparation), which employs a three-step analysis. First, the genome was divided into fixed-base windows and the average coverage depth was calculated for each window. The window size was set to be 100bp in this study. The relative coverage depth was defined as the ratio between the average window coverage and the median of the average window coverage on a set of reference chromosomes that have no gross CNVs based on chromosome-by-chromosome paired tumour/normal coverage analysis. The difference of the relative coverage depth between the tumour sample and its matching normal sample was corrected for the GC content of the window and used as the signal for calling CNVs. Second, each chromosome was segmented using a recursive partitioning method on the difference of the tumour versus normal signal. Third, the segments were merged to ensure a genome-wide error rate not greater than 0.05. CNVs were manually reviewed by comparison with Affymetrix SNP 6.0 or 500K CNV results and structural variation breakpoints identified by CREST. Missing breakpoints that define the CNV boundaries were manually mapped at base-pair resolution by visual inspection of the soft-clipped reads in the immediate neighbourhood of the predicted CNV boundaries. All analyses were performed in R[69] (64-bit version 2.9.1, with basic and tree package, version 1.0-28).

In addition to CONSERTING, all samples were analysed by CNV-Seq[70] at SJCRH for the purposes of comparison. For CNV-Seq, uniquely mapped reads with mapping quality >= 35 were used as the input to calculate the theoretical minimum window size according to a preset *P* threshold of 0.001 and $\log_2$ copy number ratio of 0.5 for each pair of tumour and normal samples. For each window, the number of read count was replaced with the mean coverage of the sample if it is less than that number before global normalisation and calculations of the $\log_2$ ratio of tumour vs. normal and the *P* value. CBS[71,72] was used to segment the $\log_2$ ratio values per chromosome and to identify candidate gain and loss regions using the following cut-offs: abs(seg.mean)>=0.5; >=8 markers per segment; and median CNV-seq *P* value for a segment <= 0.001. Finally, the filtered segments were merged where the inter-segment distance is less than 500kb and copy number difference < 0.25.

One of the samples, SJTALL007, was also analysed by SegSeq[73] using the default parameters and with the local window size set to 400. The input for SegSeq was generated by recording the genomic positions of the first base of each read in tumour and the normal bam files. CNVs less than 75 kb or with copy number ratio (tumour/normal) change less than 20% were filtered. For SJTALL007, the initial output from SegSeq includes a total of 98,878 CNV segments while the filtered result has 2,848. Only 18 CNV segments were identified by SNP array analysis of in this data set.

Identification of loss-of-heterozygosity

Regions of loss-of-heterozygosity (LOH) were identified from the high quality single nucleotide variants (SNVs). First, heterozygous SNVs with mutant allele frequency between 40-60% in the germline sample were used to estimate the LOH signal. For each heterozygous SNV, the LOH signal was calculated as the absolute mutant allele frequency difference between the tumour sample and the germline sample. Second, chromosomes were segmented and segments were merged on the LOH signal by the methods described in the *Identification of copy number variations* section.

Assessment of Telomere Length

The total number of telomeric reads were assessed by searching the next generation sequencing .bam file for reads containing the repetitive telomeric motif $(TTAGGG)_4$ (ref. [74]). The total numbers of reads were then normalised to the average genomic coverage.

SIFT/PolyPhen2 Analysis

Predicting deleterious effects of amino acid substitutions (AAS) on protein function is valuable for variant prioritization. Several open source programs are available for amino acid substitution effect prediction. SIFT[75] and PolyPhen2 (ref. [76]) were selected due to their availability for download to perform analyses locally. The SIFT pipeline was modified to query a local NR database. During preliminary analyses, SIFT was found to be highly dependent on the sequence database selected. The NR (March 2010) database was chosen for SIFT analysis because results using NR corresponded more closely with results from the online version of PolyPhen2. UNIREF100 (ref. [77]) was used for PolyPhen2 sequence queries.

RNA-Seq analysis

All Illumina paired-end reads were aligned to the following 4 database files using BWA (0.5.5) aligner: (1) human NCBI Build 36 reference sequence; (2) RefSeq; (3) Sequence file that represents all possible combinations of non-sequential pairs of RefSeq exons; (4) AceView flat file downloaded from UCSC which represents transcripts constructed from human EST. The mapping results from (2) to (4) were mapped to the human reference genome coordinates. The final BAM file was constructed by selecting the best alignment in the four databases. Coverage, SNV, indel and SV analyses were carried out by SJCHR using the methods described above. One novel indel in *EVX1* was detected by RNA-Seq, but not whole genome DNA sequencing because the site has low sequence coverage in whole-genome sequencing. Subsequent validation by Sanger sequencing in both tumour and normal germline DNA showed it to be a germline variant.

Pathway analysis - enrichment analysis of lesion data by a genomic random interval model (GRIN)

A genomic random interval (GRIN) model was used to evaluate the statistical significance of the number of subjects the number of times a somatic genomic abnormality (copy number alteration, indel, mutation, or structural alteration) overlapped a gene belonging to a pre-defined gene-set. Under the null hypothesis, the GRIN model assumes that a genomic abnormality of length L base pairs occurs randomly along any interval locus along a chromosome of length K base pairs from (1,L) to (K-L+1,K) with equal probability. Under this model, the probability that an abnormality of length L base pairs overlaps a gene beginning at position A and ending at position B on a chromosome of length K is

$$P(L,K,A,B) = (\min(K-L+1,B)-\max(0,A-L))/(K-L+1).$$

This probability represents the proportion of possible starting loci for an interval of length L on a chromosome of length K such that the interval overlaps the gene interval (A,B). An example is shown in Supplementary Figure 1. Note that this probability increases in lesion size L and gene size B-A and accounts for gene location (A,B) and chromosome size K. In a similar way, we derive the null probability that a random interval of length L overlaps g = 0,1,2,…,G of a set of G genes with loci (A_1,B_1), (A_2,B_2), …, (A_G,B_G) by determining the proportion of possible start loci for a random interval of length L that overlap g=0,1,2,…,G genes.

For a set of lesions on the same chromosome, GRIN derives the probability that each lesion impacts a given number of genes belonging to the pre-defined gene-set as described above. The lesions are assumed to be independent and the probability distribution for the total number of gene-lesion overlap events on a chromosome is the convolution of the individual lesions' probability distributions. The null distribution for the total number of gene-lesion overlaps for a subject is the convolution of the individual chromosomes' distributions. The null distribution for the number of lesion-gene overlaps across a cohort is the convolution of the individuals' distributions and the null distribution for the number of subjects with at least one lesion-gene overlap is the convolution of the individual-specific Bernoulli distributions with success probability defined as observing at least one lesion-gene overlap.

In most cases, the null distributions described above are computed exactly. In some complex cases, the distributions are estimated by Monte Carlo simulation of the loci of the lesions. One-sided hypothesis tests are performed by computing the probability that the number of lesion-gene overlaps is greater than or equal to the observed number. Similarly, a one-sided test is performed for the number of subjects in a cohort with at least one lesion-gene overlap.

Exome capture and sequencing

Tumour and normal DNA for samples SJTALL169, 192 and 208 was subjected to exome capture using the SureSelect Human All Exon 50Mb XT kit protocol version 1.1 according to the manufacturer's protocol (Agilent, Santa Clara, CA). These samples were selected for sequencing based on the criteria of being (1) ETP T-ALL, and (2) lacking lesions targeting lymphoid development and/or cytokine receptor / Ras signalling following the first phase of recurrence mutation testing of 23 genes (*BCL11B, BRAF, DCLRE1C, DNM2, ECT2L, EP300, FBXW7, FLT3, GATA3, HISTH1B, HNRPA1, HNRPR, IL7R, KRAS, JAK1, JAK3, NOTCH1, NRAS, PHF6, PTEN, RELN, RUNX1* and *SMARCA*).

Briefly, 3 micrograms of DNA were sheared by acoustic fragmentation using a Covaris E210 and purified using AMPure XP beads (Beckman Coulter Genomics). The quality of the fragmentation and purification was assessed using the Agilent 2100 Bioanalyzer. The fragment ends were repaired, "A" bases were added to the 3' end of the DNA fragments and an indexing-specific paired-end adapter was ligated to the fragments. The adapter-ligated library was amplified using 6 cycles of PCR and the quality, quantity and size distribution of the PCR products were assessed using the Agilent 2100 Bioanalyzer. 500 nanograms of the sample library were hybridised with the biotinylated RNA library for 24 hours at 65ºC. Bound DNA was purified using streptavidin coated magnetic beads (Life Technologies) and subjected to stringency washes. The captured library was amplified and index tag #3 was added using 12 cycles of PCR. The quality, quantity and size range of the library was assessed using the Agilent 2100 Bioanalyzer. The resulting libraries were quantitated with the QPCR NGS library quantification kit (Agilent Technologies) using a PhiX control library (Illumina) as an external standard. Clustered flow cells were generated on a Cbot (Illumina) using the TruSeq PE Cluster Kit version 2 (Illumina) with 6 picomoles of exon enriched library. The flow cells were loaded onto an Illumina GAIIX for a paired end 2x101 cycle sequencing run using SCS version 2.8 software and SBS version 5 reagents. Three lanes were sequenced for each of the diagnosis and matched remission sample. The resulting base call files were converted to the FASTQ format using CASAVA version 1.8.1. SNVs, indels and SVs were detected according to the methods described for the analysis of WGS data by SJCRH as described above. Coverage metrics are shown in Supplementary Table 8.

Structural modelling of EZH2 mutations

Protein structures were obtained from the Protein Databank (PDB) (www.pdb.org) (July 2011 release)[78]. Limited structural information is available for the EZH2 protein and a homology model was generated based on previous studies[34]. Briefly, amongst protein sequences in the PDB, the sequence of EZH2 showed greatest similarity to that of the SET domain of MLL1 (36% identity between residues 599-732 (isoform A) of EZH2 and residues 3816-3950 of MLL1) (Supplementary Figure 17). The noted domain of EZH2 was threaded into the structure of the MLL1 SET domain (PDB: 2W5Z)[79] using the automatic method in Swiss-Model to generate a homology model[80]. Also included in the model are S-Adenosylhomocysteine, the product of methyl transfer from S-Adenosylmethionine, and a dimethylated lysine substrate peptide, based upon their positions in the MLL1 SET domain structure. The model was judged of suitable quality based on the following criteria: acceptable Ramachandran angles, no unfavorable steric clashes, and overall structural agreement with experimentally determined SET domains (PDB: 3OPE[81], 3K5K[82], 3HNA[83], and 2W5Z[79]). Furthermore, conserved SET domain residues are located in similar structural positions in the homology model and the reference MLL1 SET domain structure[79]. Residues (atoms) defined as interacting are those within 4 Å of one another. Mutations and graphics were generated using PyMOL[84] and Espript[85].

Gene expression profiling

To examine the level of expression of genes targeted by recurring sequence mutation, we examined microarray-based gene expression profiling data of B-progenitor and T-lineage ALL and normal haemopoietic cell samples generated using Affymetrix U133A microarrays (Affymetrix, Santa Clara, CA) according to the manufacturer's instructions, with data processed using Microarray Suite 5.0 (Affymetrix) as previously described[86]. This cohort comprised 575 samples including CD10$^+$CD19$^+$ normal B cells (N=4); CD34$^+$ bone marrow cells (N=4); *ETV6-RUNX1* (N=99); hyperdiploid ALL with greater than 50 chromosomes (N=116); hypodiploid (N=23); B-ALL with normal or miscellaneous karyotype (N=153); *MLL*-rearranged ALL (N=30); *BCR-ABL1* positive ALL (N=23); T-lineage ALL (N=83); and *TCF3-PBX1* positive ALL (N=40)(NCBI gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/) accession GSE33315).

To examine pathway dysregulation in ETP ALL, we performed gene expression profiling of 12 ETP and 40 non-ETP T-lineage ALL samples using Affymetrix GeneChip HT HG-U133+ PM arrays (Supplementary Table 6; GEO accession GSE28703). Statistical analyses were performed using R[69], Bioconductor version 2.6 (ref. [87]) and Spotfire Decision Site 9.1.1 (Tibco, Somerville, MA). For HT U133+ arrays, all samples were normalised by the RMA algorithm. Probesets that did not pass the background signal threshold (twice the average signal on the control probes with different GC content) across all samples were excluded for differential expression analysis, which was performed using *limma*[88] and estimation of false discovery rate[89]. For Affymetrix U133A arrays (GEO accession GSE28497), all samples were normalised to target intensity 500 in the MAS 5 algorithm. Probe sets with absent calls for all samples were excluded. Gene Set Enrichment Analysis (*GSEA-P*[90,91]) and the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (refs. [92,93]) were used to assess pathway enrichment. For GSEA, we used gene sets obtained from the Molecular Signatures Database, and gene sets of normal human haemopoietic progenitors, leukaemia stem cells and B-progenitor ALL. These included the normal human haemopoietic stem cell (Lin- CD34+ CD38- CD45RA- CD90+ CD49f+), granulocyte macrophage precursor (GMP; Lin- CD34+ CD38+ CD7- CD10- CD135+ CD45RA+) and human early T cell precursor (CD34+ CD1a- cells isolated from neonatal thymi)(refs. [38,39,94] and J.E.D. *et al.*, unpublished data). The signature of leukaemia stem cells in acute myeloid leukaemia was obtained from Eppert et al..[40] The gene expression profile of high-risk B-progenitor childhood ALL cases predicted to be at high risk of relapse was obtained from Mullighan *et al*..[41]

Reconstruction of the transcriptional network of ETP ALL

We used the ARACNE algorithm[37,95] to reconstruct gene networks that are differentially expressed in ETP versus non-ETP T-ALL tumours. Expression profiles of 40 non-ETP and 12 ETP tumour samples were characterized using Affymetrix HU133 PM Plus 2.0 microarray and normalised using the RMA algorithm. A total of 12,789 probe sets in 7,251 genes have differential expression in these two subgroups (*P*<0.05 in *limma* analysis). Using these probe sets, a consensus bootstrapping network was built based on 100 bootstrap step with *P*<1e-5 and DPI tolerance 0.1. Transcription networks (e.g. the regulons) were inferred by using only genes annotated as transcription factors with a more stringent significance threshold (i.e. *P*<1e-7). For each regulon, a principal component analysis was run using a data matrix of (m samples x n probe sets). Spearman rank correlation between the phenotype (ETP versus non-ETP) and the first principal component on the sample axis (m values) was calculated to assess the significance of association between each regulon and ETP/non-ETP status.

Transformation assays of IL7R mutants in Ba/F3 cells

The IL7R LL242-243>DTRVYNSICL, SLILIVPCACELinsA254, IL241-242TC, I241>ITLYCKT, LL242-243>SPCI, V253>GFSV and GCinsL243 mutations were introduced into MSCV-mIL7R-IRES-GFP and MSCV-mIL7R-IRES-hCD4 retroviral plasmids using the Quikchange II XL kit (Stratagene, Santa Clara, CA) as previously described[96]. Production of ecotropic retroviral supernatants, transduction of murine haemopoietic Ba/F3 cells, cytokine withdrawal and proliferation assays were performed as previously described[96]. Briefly, vectors were packaged into replication-incompetent, ecotropic retroviral particles by the triple plasmid (pMD gagpol and pCAG4-Eco) system. Murine pro-B Ba/F3 and P2RY8-CRLF2-GFP-expressing Ba/F3 cells[23] were transduced with wild-type or mutant Il7r retroviral supernatants. Transduced cells were maintained in RPMI-1640 with 10% FCS, penicillin-streptomycin, and L-glutamine. To assess growth factor independence, cells were washed 3 times and were plated at 500,000 cells per millilitre in media without cytokine. Growth was monitored daily by using a ViCell cell counter (Beckman Coulter, Danvers, MA).

Transformation assays of IL7R mutants in the MOHITO T-cell line

The mouse cytokine dependent leukaemic T-cell line MOHITO was recently established from a BALB/c mouse, which had spontaneously developed a T-ALL like disease[25]. Cells were cultured in RPMI-1640 (Life Technologies) supplemented with 20% fetal calf serum containing 5 ng/ml IL-2 and 10 ng/ml IL-7 (Peprotech).

Viral production and retroviral infection of MOHITO cells was performed as described previously with minor modifications[25,97]. Briefly, non-tissue culture treated 6-well plates were coated with RetroNectin solution overnight in the fridge (Takara Bio Inc.) and blocked with 0.5% FBS in PBS for 30 minutes before use. Viral supernatant was pre-loaded onto coated plates by centrifugation (1000 g, 120 minutes, 30°C). After centrifugation, viral supernatant was removed, plates were washed with 2 ml PBS and cells were added at a density of $0.5x10^6$ cells/ml. Retroviral transduction was achieved using standard spin-infection procedure (2000g, 60 minutes, 30°C). Cells were placed in an incubator for 72 hours to recover before determination of transduction efficiency and performance of subsequent experiments. Cells were split 24 hours before transduction was performed to achieve exponential growth.

For transformation assays, MOHITO cells were washed twice in PBS to ensure complete removal of cytokines. After the last wash step cells were resuspended in cytokine-free culture media at a cell density of 0.3 x 10^6 cells/ml. 200 μl of prepared cell suspensions were seeded out in 96-well plates (*n*=3). The number of GFP-positive cells was determined at day of seed-out and at indicated time points afterwards by flow cytometry. Ectopic expression of BCR-ABL1 transforms MOHITO cells rapidly to cytokine independence and was included as a positive control[25]. Western blotting of MOHITO cells expressing Il7r mutant alleles was performed using 30μg of whole cell lysate prepared using NuPAGE LDS buffer (Life) electrophoresed through 4-12% NuPAGE Bis-TRIS gels that after transfer were probed with mouse Il7ra (CD127) antibody (R&D Systems) and beta-actin (AC-15, Sigma Aldrich) under reducing (dithiothreitol or beta-mercaptoethanol) and non-reducing conditions.

Phosphoflow analysis of MOHITO cells

For analysis of phosphosignalling, MOHITO cells were starved overnight (RPMI/0.5% BSA) or left in serum, treated with or without 3 μM JAK inhibitor 1 (EMD Biosciences) for 1 hr. MIG and WT-Il7r cells were stimulated with recombinant IL-7 at 10ng/ml for 15 minutes before fixation and permeabilisation. Cells were stained with anti-pSTAT5 (Y694; Cell Signaling Technology) and Alexa-Fluor 647 conjugated anti-rabbit IgG secondary antibody (Life Technologies). The samples were collected on a FACSCalibur (BD Biosciences) using Cell Quest software (BD Biosciences), and analysed with FlowJo (Tree Star).

Lineage-negative enrichment and colony assays of murine haemopoietic cells

Experiments were approved by the St Jude Children's Research Hospital Institutional Animal Care and Use Committee. Bone marrow mononuclear cells (BMMC) were harvested from 8-wk-old wild-type (WT) or Arf$^{-/-}$ C57BL/6 mice and labelled with biotin-conjugated lineage antibodies (Ly-6G, CD11b, CD45R, CD5, TER-119; PharMingen), followed by incubation with streptavidin-coated magnetic beads (Dynabeads M-280 Streptavidin, Dynal). Lineage-negative cells were purified by magnetic separation and cultured for 48 hr in IMDM/20% FCS supplemented with penicillin-streptomycin, L-glutamine, recombinant mouse IL-3 (10 ng/ml), IL-6 (20 ng/ml), IL-7 (10ng/ml), Flt-3 ligand (40ng/ml) and stem cell factor (SCF; 50 ng/ml) (Peprotech). Cells were infected on RetroNectin-coated plates for 48 hr (Takara Bio Inc.) with MSCV-IRES-GFP retrovirus expressing WT or mutant Il7r (IL241-242TC, LL242-243>SPCI, GCinsL243, V253>GFSV). Transduced GFP$^+$ cells were obtained by fluorescence-activated cell sorting. For clonogenic assays, 10,000 cells were plated in duplicate in Methocult M3231 (Stem Cell Technologies, Inc., Vancouver, BC, Canada) with the appropriate factors (SCF, Flt-3 ligand, IL-7) and colonies were scored 7 days later. For re-plating, 10,000 cells were cultured in identical conditions, with colonies counted on day 12.

Phosphoflow analysis of primary human T-ALL leukaemic cells

Patient T-ALL leukaemia samples from the Children's Oncology Group tissue bank were thawed, washed, and adjusted to 2x10^6 cells/ml in serum free media containing 0.5% BSA. JURKAT, a non-ETP T-cell ALL sample, and normal human thymocytes (obtained with informed consent from children undergoing cardiac surgery) were included as controls. After resting cells for one hour at 37°C, 0.5x10^6 cells were plated and stimulated with 200 μM pervanadate for 20 minutes. Cells were subsequently fixed, permeabilised, and stained with antibodies specific for CD3-PECy7 (BD), CD7-PECy5 (eBioscience), Caspase-3-V450 (BD), phospho-Stat5-APC (BD), phospho-AKT-PE (Cell Signaling), phospho-S6-APC (BD) and phospho-ERK-PE (BD) as previously described[98]. Samples were analysed on an LSRII flow cytometer (BD Biosciences). Data were analysed using Cytobank (Stanford University) and FlowJo 8.8.2 (Tree Star) software.

Analysis of genetic alterations and outcome

Outcome data were available for 102 of the 106 patients examined. Associations between genetic alterations and treatment outcome (induction failure, event free survival and relapse) were performed as previously described[41,99-102]. Analyses were performed using SAS (SAS v9.1.2, SAS Institute, Cary, NC) and SPLUS (SPLUS 7.0, Insightful Corp., Palo Alto, CA) and StatXact (v 8.0.0, Cytel Inc, Cambridge, MA). Univariable and multivariable analyses considering genetic lesions, ETP status and age were performed. Presentation peripheral blood leukocyte count and minimal residual disease (MRD) were not considered as these data were not available for several cohorts. Genetic variables examined include all recurring sequence and structural genetic alterations studied in T-ALL tabulated in Supplementary Table 18.

METHODS REFERENCES

44 Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).

45 Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-1066 (2009).

46 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

47 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

48 Edmonson, M. N. *et al.* Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* **27**, 865-866 (2011).

49 Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-36 (2009).

50 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).

51 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

52 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).

53 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).

54 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

55 Huang, X. Q. & Miller, W. A Time-Efficient, Linear-Space Local Similarity Algorithm. *Advances in Applied Mathematics* **12**, 337-357 (1991).

56 Flicek, P. *et al.* Ensembl 2011. *Nucleic Acids Research* **39**, D800-D806 (2011).

57 Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res* **36**, D25-30 (2008).

58 Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).

59 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).

60 Mullighan, C. G. *et al.* CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature* **471**, 235-239 (2011).

61 Zhang, J. *et al.* SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS Comput Biol* **1**, e53 (2005).

62 Chen, K. *et al.* PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res* **17**, 659-666 (2007).

63 Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, 355-358 (2004).

64 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681 (2009).

65 Sindi, S., Helman, E., Bashir, A. & Raphael, B. J. A geometric approach for classification and comparison of structural variants. *Bioinformatics* **25**, i222-230 (2009).

66 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).

67 Rozen, S. & Skaletsky, H. J. in *Bioinformatics Methods and Protocols: Methods in Molecular Biology* eds S. Krawetz & S. Misener) 365-386 (Humana Press, 2000).

68 Zhang, Q. *et al.* CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics* **26**, 464-469 (2010).

69 *R Development Core Team. R: A language and environment for statistical computing*, http://www.R-project.org> (2009).

70 Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).

71 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004).

72 Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-663 (2007).

73 Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99-103 (2009).

74 Castle, J. C. *et al.* DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics* **11**, 244 (2010).

75 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

76 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).

77 Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288 (2007).

78 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

79 Southall, S. M., Wong, P. S., Odho, Z., Roe, S. M. & Wilson, J. R. Structural basis for the requirement of additional factors for MLL1 SET domain activity and recognition of epigenetic marks. *Molecular cell* **33**, 181-191 (2009).

80 Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L. & Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research* **37**, D387-392 (2009).

81 An, S., Yeo, K. J., Jeon, Y. H. & Song, J. *Structural Basis of Auto-inhibitory mechanism of Histone methyltransferase*, http://www.pdb.org/pdb/explore/explore.do?structureId=3OPE

82 Liu, F. *et al.* Discovery of a 2,4-diamino-7-aminoalkoxyquinazoline as a potent and selective inhibitor of histone lysine methyltransferase G9a. *Journal of medicinal chemistry* **52**, 7950-7953 (2009).

83 Wu, H. *et al.* Structural biology of human H3K9 methyltransferases. *PLoS One* **5**, e8570 (2010).

84 The PyMOL Molecular Graphics System. v.1.3 (San Carlos, CA, 2011).

85 Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **15**, 305-308 (1999).

86 Ross, M. E. *et al.* Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **102**, 2951-2959 (2003).

87 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).

88 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).

89 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**, 289-300 (1995).

90 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

91 Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251-3253 (2007).

92 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

93 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).

94 Doulatov, S. *et al.* Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nature immunology* **11**, 585-593 (2010).

95 Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).

96 Mullighan, C. G. *et al.* JAK mutations in high-risk childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **106**, 9414-9418 (2009).

97 De Keersmaecker, K. *et al.* Fusion of EML1 to ABL1 in T-cell acute lymphoblastic leukemia with cryptic t(9;14)(q34;q32). *Blood* **105**, 4849-4852 (2005).

98 Kotecha, N. *et al.* Single-cell profiling identifies aberrant STAT5 activation in myeloid malignancies with specific clinical and biologic correlates. *Cancer Cell* **14**, 335-343 (2008).

99 Peto, R. *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *Br J Cancer* **35**, 1-39 (1977).

100 Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* **50**, 163-170 (1966).

101 Gray, R. J. A class of *K*-sample tests for comparing the cumulative incidence of a competing risk. *Annals Statistics* **16**, 1141-1154 (1988).

102 Fine, J. P. & Gray, R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* **94**, 496-509 (1999).