Legends for supplementary figures and tables:

Figure S1: Expression of the engineered ETV6-RUNX1 fusion protein. (A) In consonance with findings for other fusion oncogenes [Brian Huntly, personal communication], we were unable to detect the presence of the *ETV6-RUNX1* fusion protein in primary hematopoetic tissue extracts from *Etv6*$^{+/RUNX11}$ mice (data not shown), so to demonstrate expression of our engineered fusion protein we modified the *pAML-IRES-SB-puro* targeting construct to add a Flag tag to the end of *RUNX11* sequence (by cloning in a PCR fragment generated with FWD primer: 5'-GCT CGC CGC CGC GCA TCC T-3' and REV primer: 5'-GGC CTT AAT TAA TCA CTT GTC GTC ATC GTC CT-3') and exchange the *Sleeping Beauty* (SB) for EGFP (by recombineering). **(B)** After targeting this construct into ES cells (*Etv6*$^{+/RUNX1\text{-}Flag}$), the Flag-tagged fusion protein was purified using immunoprecipitation of whole cell lysates. Western blotting using an anti-Flag antibody showed two bands, ~100 and ~110 KDa, representing translation starting from M1 and M43 alternative start codons. A Flag-tagged human *ETV6-RUNX1* cDNA transiently transfected into ES cells was used as a positive control for Western blotting, producing an ~110 kDa band.

Figure S2: Characterisation of the *Etv6*$^{+/RUNX1}$ allele *in vivo*. A. Intercrossing of *Etv6*$^{+/RUNX1}$ mice (on a mixed 129;C57 background) resulted in homozygotes (*Etv6*$^{RUNX1/RUNX1}$ mice) only seen prior to embryonic day 10.5 (E10.5; representative images of E10.5 embryos from a *Etv6*$^{+/RUNX1}$ intercross shown). **B.** Penetrance of the *Etv6*$^{+/RUNX1}$ allele is modified by genetic background. The *Etv6*$^{+/RUNX1}$ allele was generated in E14Tg2a ES cells (129P2Ola/Hsd genetic background) and the resulting chimaeras bred with either 129S5SvEv^{Brd} or C57BL/6J wildtype mice for germline transmission. When the offspring (either on a *129* or mixed *129;C57* background) were bred with a C57BL/6J wildtype mouse, sub-Mendelian penetrance of the *Etv6*$^{+/RUNX1}$ allele was observed (Chi-squared Test, 2-tailed; *P* = 0.0035). In contrast, sub-Mendelian penetrance was not observed (taking into account embryonic lethality of homozygous allele) when *Etv6*$^{+/RUNX1}$ mice on a mixed 129;C57 background were intercrossed.

Figure S3: Immunophenotyping hematological cells of the bone marrow from *Etv6*$^{+/+}$ and *Etv6*$^{+/RUNX1}$ mice. Bone marrow from *Etv6*$^{+/+}$ and *Etv6*$^{+/RUNX1}$ mice at 3, 6 and 12 months of age was analysed by immunophenotyping on the flow cytometer using myeloid (Mac-1, Gr-1), T-cell (CD4) and B-cell (B220) markers. Results expressed as

percentage of the total cell population of the bone marrow (mean ± SD, n=4-6 mice at each timepoint).

Figure S4: Immunophenotyping of the B220+ bone marrow cells from *Etv6*$^{\mathit{+/RUNX1}}$ mice with BCP-ALL. FACS plots from the bone marrow of representative mice demonstrate B220+ cells showing a Mac1^{+}, CD43^{+}, AA4.1^{+}, CD24^{+}, CD4^{-}, CD19^{-}, IgM^{-}, IgD^{-}, BP1^{-}, IL7Ra^{-} phenotype.

Figure S5: A breakdown of the types of reads. Leukemic genomic DNA from 73 *Etv6*$^{\mathit{+/RUNX1}}$*; T2Onc*$^{\mathit{+/Tg}}$ (*EROnc*) mice were used in a ligation-mediated PCR method to produce barcoded PCR products that were pooled and sequenced on the 454 GS FLX System to generate 695,504 reads. After pre-processing the reads and aligning them to the mouse genome there were 341,320 uniquely aligned sequences (representing 51.5% of the reads; shown in dark blue). The remaining 48.5% of the reads were discarded due to the reasons listed in the key.

Table S1. Gene expression analysis of *ETV6-RUNX1* BCP-ALL cases. Gene expression data were obtained using splenic RNA from 15 *ETV6-RUNX1* BCP-ALL cases and 3 *ETV6-RUNX1* non-diseased mice and analysed on Illumina Mouse WG-6 v2.0 beadchips. Data were analysed and p-value adjusted (as described in the Methods) to yield a sorted list of differentially expressed genes. The differentially expressed genes with an adjusted p value <0.05 are listed. Differential expression levels are shown as log fold change (LogFC) in the BCP-ALL samples relative to the controls (a negative value indicates decreased expression).

Table S2. Gene set enrichment analysis (GSEA). GSEA was performed in which the genes with differential expression were sorted on the basis of gene function using Ingenuity Pathway Analysis software. The top 5 canonical pathways identified by the program are shown. The ratio shows the number of differentially expressed genes as a proportion of the total number of genes in that pathway.

Table S3. Gaussian Kernel Convolution (GKC) common insertion sites (CIS) identified in all the samples (n=71 leukaemias). *Total number of insertions in the gene (this may be higher than the number of ‘unique’ insertions as some samples had multiple insertions in the same gene at different sites).