SUPPLEMENTARY INFORMATION FOR

The genetic basis of early T-cell precursor acute lymphoblastic leukaemia

Contents

SUPPLEMENTARY RESULTS 5

Mutation rate and spectrum in ETP ALL 5

Structural Variants and analysis of SV breakpoint sequences 5

Chromosomal rearrangements in ETP ALL 6

Deleterious mutations in ETP ALL 7

Ras pathway and signalling mutations in ETP ALL. 7

Inherited variants in *IL7R* and the risk of T-ALL 8

Mutations affecting lymphoid development in T-ALL 8

Analysis of sequence variations in matched non-tumour samples 9

Structural modelling of EZH2 mutations identified in T-ALL 9

SUPPLEMENTARY TABLES 13

SUPPLEMENTARY FIGURES 41

SUPPLEMENTARY REFERENCES 78

SUPPLEMENTARY TABLES

Supplementary Table 1. Details of whole genome sequenced cases. 13

Supplementary Table 2. DNA copy number alterations in ETP and non-ETP T-ALL identified by SNP array analysis 14

Supplementary Table 3. Cumulative DNA copy number alteration and lesion number in ETP and non-ETP T-ALL identified by SNP array analysis 14

Supplementary Table 4. Results of DNA copy number alteration for ETP and non-ETP ALL using GISTIC 14

Supplementary Table 5. Primer sequences for RT-PCR of chimeric fusion transcripts and rearrangements in ETP ALL. 15

Supplementary Table 6. Cases studied by gene expression profiling. 16

Supplementary Table 7. Coverage data of cases sequenced 17

Supplementary Table 8. Coverage data for exome sequencing 18

Supplementary Table 9. Summary of somatic sequence mutations, somatic structural variations and somatic copy number alterations identified in the 12 ETP ALL cases. 19

Supplementary Table 10. Comparison of mutation frequencies to previously published whole genome sequencing studies in cancer 20

Supplementary Table 11. Validated sequence mutations (including both substitution variants and indels) in RefSeq genes for WGS cases. 21

Supplementary Table 12**.** Numbers of sequence mutations in Refseq genes for WGS cases. 21

Supplementary Table 13. Validated structural variations found in the 12 WGS cases. 23
Supplementary Table 14. Copy number alterations identified by analysis of WGS and SNP microarray data for the 12 discovery WGS ETP ALL cases. 25
Supplementary Table 15. Correlation between abnormalities detected by cytogenetics and DNA copy number alteration / structural variant analysis of WGS data. 26
Supplementary Table 16. Chimaeric fusion genes predicted by CREST analysis of whole genome sequencing and RNA-seq data. 27
Supplementary Table 17. List of all non-silent validated mutations from WGS (n=12) and recurrence screening (n=94). 28
Supplementary Table 18. Matrix showing copy number alterations and sequence mutations for all cases. 28
Supplementary Table 19. Frequency of recurring somatic genetic alterations in ETP and non-ETP T-ALL. 29
Supplementary Table 20. Frequency of cytokine/Ras signalling and developmental mutations in ETP and non-ETP T-ALL 30
Supplementary Table 21. Lack of association between the IL7R rs6897932 (T244I) genotype and the risk of T-ALL. 31
Supplementary Table 22. Sequence variants identified on analysis of WGS data obtained from matched non-tumour samples 32
Supplementary Table 23. Listing of pathway analysis results for the WGS cases 33
Supplementary Table 24. *limma* gene expression signature of 12 ETP versus 40 non-ETP T-ALL samples. 34
Supplementary Table 25. Reverse engineering of transcription factor networks in ETP ALL using ARACNE. 35
Supplementary Table 26. Associations between ETP status, genetic lesions and outcome in T-ALL. 36
Supplementary Table 27. Prognostic effect of genetic alterations after adjusting for ETP status. CI, confidence interval. HR, hazard ratio. 39
Supplementary Table 28. Comparison of the spectrum of genetic alterations in ALL and AML. 40

SUPPLEMENTARY FIGURES

Supplementary Figure 1. The genomic random interval (GRIN) model 41
Supplementary Figure 2. Number and burden of DNA copy number alterations in ETP and non-ETP ALL 41
Supplementary Figure 3. Regions of significant DNA copy number alteration (GISTIC). 42
Supplementary Figure 4. Genome coverage of WGS cases 44
Supplementary Figure 5. The mutation spectrum of ETP ALL. 45
Supplementary Figure 6. CIRCOS plots of genetic alterations in all 12 WGS cases 47

Supplementary Figure 7. Telomere shortening in ETP ALL....49
Supplementary Figure 8. An example of multiple complex inter-chromosomal rearrangements in sample SJTALL002. 50
Supplementary Figure 9. Complex rearrangement of *ETV6* in SJTALL002. 51
Supplementary Figure 10. Complex rearrangements in case SJTALL012. 53
Supplementary Figure 11. RNA-seq identifies RUNX1 rearrangement in SJTALL012. 54
Supplementary Figure 12. Reciprocal inter-chromosomal translocations in ETP ALL. 55
Supplementary Figure 13. Detection of *ETV6-INO80D* in case SJTALL208 by whole exome sequencing. 56
Supplementary Figure 14. *ETV6* deletions and sequence mutations in ETP ALL. 57
Supplementary Figure 15. Protein domain and mutation plots for targets of recurring and novel sequence mutation in T-ALL. 58
Supplementary Figure 16. Deletions and mutations of *IKZF1* (IKAROS) in ETP ALL. 62
Supplementary Figure 17. Polycomb repressor 2 complex mutations in ETP T-ALL 63
Supplementary Figure 18. Sequence alignment of the EZH2 SET domain with the sequence of the structurally characterised MLL1 SET domain (PDB: 2W5Z). 64
Supplementary Figure 19. Structural modelling of EZH2 mutations. 64
Supplementary Figure 20. *SETD2* alterations in ETP ALL 68
Supplementary Figure 21. Frequency and association lesion matrix. 69
Supplementary Figure 22. Affymetrix U133A data showing box plots of expression levels of genes targeted by recurring sequence alterations in T-ALL. 70
Supplementary Figure 23. Gene expression profiling analysis of ETP ALL. 72
Supplementary Figure 24. Unsupervised hierarchical clustering of T-ALL showing recurring genetic alterations 73
Supplementary Figure 25. DAVID analyses of dysregulated pathway expression in ETP ALL ..74
Supplementary Figure 26. Phosphosignalling analysis of ETP and non-ETP T-ALL. 76
Supplementary Figure 27. Gene set enrichment analysis of ETP ALL 77

SUPPLEMENTARY RESULTS

Mutation rate and spectrum in ETP ALL

The estimated mean mutation rate was 3.08×10^{-7} per base (range 1.03×10^{-7} to 8.23×10^{-7}), which is 3-4-fold lower than mutation rates reported in adult tumours, with the exception of acute myeloid leukaemia (Supplementary Table 10). C>T/G>A transition was the most common change in all cases, accounting for up to 50% of all somatic sequence mutations, suggesting no exposure to environmental carcinogens as observed in adult cancers[1]. A total of 153 validated somatic sequence variants were found to cause amino acid changes in Refseq[2] genes (Supplementary Table 11) with a mean of 8.5 (range 3-16) missense, 0.67 nonsense (range 0-3) and 1.5 insertion/deletion (indel, range 0-4) mutations per case. Despite a low overall mutation rate, 54% of the missense mutations were predicted to be deleterious, suggesting that many of these variants are directly involved in leukaemogenesis (Supplementary Table 12).

Structural Variants and analysis of SV breakpoint sequences

We detected 181 SVs across the WGS cases, 80% of which were validated by Sanger sequencing (Supplementary Tables 13-14). This included 44 CTX (mean 4 per case, range 0-12), 32 ITX (mean 3, range 0-7), one INV, 53 DEL (mean 4, range 0-10) and 16 INS (mean 1, range 0-5) (Figure 1, Supplementary Figure 6; see also Supplementary Results). Most abnormalities identified by cytogenetics were also evident on analysis of WGS data (Supplementary Table 15). Two samples (SJTALL011 and SJTALL012) had a 'closed chain' pattern of chromosomal breakage and rejoining with no concomitant copy number alteration as recently described in prostatic carcinoma[3].

Analysis of the SV genomic breakpoints indicated that the majority of SVs arose from non-homologous end-joining rather than alternative mechanisms such as recombinase-activating gene (RAG) mediated recombination. In addition, the proportion of SVs with evidence of micro homology at the breakpoints was highly variable (absent in case SJTALL004 to 75% in SJTALL009). Many focal SVs in ALL are thought to arise from aberrant (Recombinase Activating Gene) RAG mediated V(D)J recombinatorial activity, as suggested by analysis of the SV breakpoints which commonly show heptamer recombination signal sequences (RSS) immediately inside the break, with the addition of non-consensus nucleotides between the SV breakpoints[4-6]. RSS consist of a conserved heptamer followed by a non-conserved spacer of 12 (RSS 12) or 23 (RSS 23) bp which is followed by a conserved nonamer. The RSS heptamer-spacer-nonamer sequences are known to be recognized by the RAG1/2 enzymes in order to

juxtapose genes to generate antigen receptors[7]. To determine if SVs were mediated by this mechanism, the following analysis was performed. 200bp sequences around break points were extracted and the RSS sites were predicted for all the sequences using the RSS database tool (http://www.itb.cnr.it/rss/index.html). This prediction tool uses a position-specific scoring matrix (PSSM) to calculate a score for each candidate site[8]. A RSS site with score greater than -38.81 and -58.45 for RSS 12 and RSS 23 respectively was considered a potential RSS site. If both break points of a SV have the RSS 12 and RSS 23 separately, the event could be mediated by this mechanism. Fifteen out of 143 SVs in this study had this feature, thus this was not the major mechanism for most of SVs. Recent data suggest that homologous sequences (e.g. Alu sequences) mediated indels and duplications are the major forms of SVs in normal individuals[9]. But for the majority of the SVs identified, the two break points do not share long homology, as the Sanger sequencing validation showed only short overlap between the mappings of fusion product and the observed break points. We next examined the fusion sequence with the break point sequences and separated the SVs into two categories: (1) micro-homology mediated SVs, in which the SV breakpoints share at least 5bp but not longer than 20bp sequence homology, and (2) non-homologous end join (NHEJ) mediated SVs. Thirteen of 143 of events have micro-homology sequence and 115 of 146 events have no homologous sequence. Thus, in this dataset, most somatic SNVs likely arise from non-homologous end-joining.

Chromosomal rearrangements in ETP ALL

Case SJTALL002 harboured a complex translocation involving chromosomes 2q33.3-34 (at *INO80D*, a component of a chromatin remodelling complex[10]), 12p13.2 (at *ETV6*, an ETS domain containing transcription factor required for definitive haemopoiesis[11] that is frequently altered in leukaemia[12-14]) and Xq13.1 (at *BG201338,* a non-Refseq EST; Supplementary Figure 9a). This rearrangement involved complex amplifications adjacent to the breakpoints at 2q33 and Xq13.1, and a focal deletion within the *ETV6* locus at 12p13.2 (Supplementary Figure 9b). This CTX resulted in the expression of three chimeric fusions: *ETV6-INO80D, INO80D-BG201338* and *BG201338-ETV6* (Supplementary Figure 9c-e). *ETV6-INO80D* is predicted to encode a 724 amino acid protein that retains the pointed domain of ETV6. *INO80D-BG201338* encodes a 653 amino acid protein of unknown function, and *BG201338-ETV6* leads to an out-of-frame fusion (Supplementary Figure 9e).

Case SJTALL012 harboured a multi-chromosomal translocation involving chromosomes 8q21.13, 7p15.2 (upstream of *HOXA13*), 21q22.12 (at *RUNX1*) and 10q22.2-24.32 (Supplementary Figure 10a) that required transcriptome sequencing in addition to WGS in order

to define the anatomy of the rearrangement. The rearrangement of 21q to 7p was predicted to juxtapose the 5' region of *RUNX1* to a non-coding gene, *NCRNA00213* (Supplementary Figure 10b). This breakpoint at 7p15.2 was located approximately 40 kb upstream of *EVX1* (even-skipped homeobox 1), and transcriptome sequencing of this case identified read-through downstream of the fusion junction, and generation of in-frame *RUNX1-EVX1* transcripts (Supplementary Figure 10c, Supplementary Figure 11).

Case SJTALL009 harboured rearrangements of chromosomes 9q34 and 5q35 resulting in expression of *NUP214-SQSTM1* and the reciprocal *SQSTM1-NUP214* transcripts (Supplementary Figure 12a-c). *NUP214* encodes a nucleoporin and is a known target of rearrangement in T-ALL[15]. *SQSTM1* encodes sequestosome 1, a multifunctional scaffold protein that binds ubiquitin and regulates activation of nuclear factor kappa-B signalling, and is mutated in Paget's disease of bone[16]. SJTALL013 harboured a rearrangement of *NAP1L1* (nucleosome assembly protein 1-like 1) at chromosomes 12q21 and *MLLT10* (AF10, a translocation partner of *MLL* in acute leukaemia) at 10p12 resulting in the expression of *NAP1L1-MLLT10* and *MLLT10-NAP1L1* (Supplementary Figure 12d-g). Case SJTALL003 had a rearrangement of *CTNNA3* (catenin, alpha 3) at 10q22.2 to *ARHGAP21* (Rho GTPase activating protein 21) at 10p12.3, but no fusion transcript was detectable by RT-PCR due to absent expression of *CTNNA3* as demonstrated by microarray gene expression profiling data for this case (data not shown).

Deleterious mutations in ETP ALL

Strikingly, comparison of the results of the mutation recurrence screening to those of the WGS cases demonstrated a marked enrichment for predicted deleterious mutations in the 212 validated non-silent mutations identified in the recurrence cohort (Supplementary Table 17). We observed a higher proportion of deleterious missense mutations (82% vs. 51%, P=0.003), indels (40% versus 12% in WGS cases, P=0.0001; and 5-6% in glioblastoma and pancreatic carcinoma genomes[17,18]) and an elevated proportion of nonsense and splice mutations expected to cause protein truncation (15% versus 8%, P=0.05), indicating that a higher proportion of the identified mutations are likely to be "driver" lesions in comparison to adult tumours.

Ras pathway and signalling mutations in ETP ALL.

Nine cases harboured *FLT3* mutations (9 ETP *v.* none in non-ETP), 16 *NRAS* (12 *v.* 4), 2 *KRAS* (both ETP), two *BRAF* (all ETP, including a G466E mutation and a focal amplification), 7 *JAK3* (all ETP), 8 *JAK1* (7 ETP, one non-ETP). The *NRAS* and *KRAS* variants were either known

activating mutations in these genes, or are located at or adjacent to these residues (Supplementary Figure 15). Similarly, the *FLT3* mutations were located in the transmembrane domain and at D835 in the tyrosine kinase domain, both of which are activating mutations in this gene. An additional case (SJTALL004) harboured a mutation in *IGF1R* (S1282I) that is located immediately distal to the kinase domain, and is predicted to be deleterious. IGF1R is expressed in a variety of tumours and signals through both the Ras/Raf/MEK/ERK and PI3K/AKT pathways[19]. IGF1R signalling has been implicated in tumour transformation and the survival of malignant cells[19], and rare mutations have been identified in solid tumours (COSMIC data at http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=gene&ln=IGF1R).

Inherited variants in *IL7R* and the risk of T-ALL

The mutational hotspot in IL7R is in close proximity to an inherited SNP, T244I, associated with the risk of multiple sclerosis[20], raising the possibility that inherited variants in IL7RA may also influence the risk of T-ALL, as has been described for other genes in ALL[21,22] or the acquisition of somatic alterations in the gene as reported for JAK2 mutations[23]. However, we found no difference between the frequency of this allele in ETP ALL or non-ETP T-ALL and a control CEPH population (Supplementary Table 21).

Mutations affecting lymphoid development in T-ALL

EP300. EP300 and the paralogous CREBBP (CBP) are transcriptional scaffolds and coactivators, histone and non-histone acetyltransferases and ubiquitin ligases that have important roles in a variety of developmental programs including haemopoietic development. Both *CREBBP* and *EP300* mutations have been recently identified in relapsed ALL[24] and lymphoma[25,26]. Existing data suggests that EP300 (p300) has an important role in lymphoid development, including T-lymphoid development. A germline mutation in the KIX domain of p300 which attenuates c-Myb and CREB binding has a strong effect on B and T cell development[27]. Conditional deletion of p300 in thymocytes using Lck-cre has a less pronounced effect on T cell development and function than does conditional deletion of *Crebbp* (Cpb)[28], but there are significant effects[29].

ETV6. ETV6 is a member of the Ets family of transcription factors. Knockout models of ETV6 loss have shown that this gene is required for definitive haemopoiesis, with Etv6-/- mice showing a profound defect in the development of multiple blood lineages, including lymphoid cells[11,30].

IKZF1. IKZF1 encodes IKAROS, the founding member of a family of zinc finger containing transcription factors expressed in multiple haemopoietic lineages. Expression of Ikaros is

required for the development of all lymphoid lineages[31], and for specification and differentiation of the T cell lineage[32]. Dominant negative mutations of *Ikzf1* are associated with the development of aggressive T-lymphoproliferative diseases in mice[33]. Mutations of *IKZF1* (either deletion or sequence mutation) are associated with aggressive B-lineage ALL (*BCR-ABL1* ALL[5] and a subtype of *BCR-ABL1* negative ALL with a transcriptional profile reminiscent of *BCR-ABL1* ALL[34,35] harbouring mutations in cytokine receptor and kinase signalling pathways[36,37]). In these "BCR-ABL1-like" B-ALL cases, mutation of *IKZF1* is associated with up regulation of haemopoietic stem cell genes, and reduced expression of B cell differentiation and signalling genes, suggesting that alteration of *IKZF1* directly contributes to a block in maturation in ALL[34]. Prior to this study, *IKZF1* alterations have been uncommonly observed in T-ALL[5,38].

RUNX1. RUNX1 forms part of the core-binding factor transcription complex, and is required for definitive haemopoiesis[39]. Runx1 is expressed in the lymphoid lineages during development, is required for the transition from the DN to DP stages of T cell differentiation in the mouse, and may directly interact with GATA3 to influence T-cell differentiation[40-42]. *RUNX1* is a common target of translocation in both ALL[12] and AML[43]. Sequence mutations of *RUNX1* have been described in myeloid malignancies[44-46], and are responsible for the familial platelet disorder with predisposition to acute leukaemia (FPD/AML, MIM 601399), an autosomal dominant disorder characterized by numerical and qualitative platelet defects and an increased risk of AML[47,48], but have not been reported in T-ALL. As in the myeloid disorders, the mutations observed in T-ALL commonly involve the Runt domain, include frameshift and nonsense mutations, and are predicted to be deleterious. Notably, two mutations were located at or near R166 (R166* and V164A) and R201 (R204*), residues known to be mutated in FPD/AML.

Analysis of sequence variations in matched non-tumour samples

Analysis of sequence variations in the matched normal DNA for each case, and identified an average of 303 (range, 183-464) non-silent variants per case, up to 50% of which were predicted to be deleterious or damaging (Supplementary Table 22).

Structural modelling of EZH2 mutations identified in T-ALL

Missense mutations were identified throughout the EZH2 protein from multiple cases of T-ALL. Several mutations clustered within the SET domain at positions Ile 646 (to Phe; I646F), Arg 679 (to His; R679H), Asn 688 (to Tyr; N688Y), and Ser 690 (to Leu; S690L). The SET domain is the catalytic domain responsible for histone N-methyltransferase activity and is therefore a common feature in most N-methyltransferase proteins[49]. Based on homology to the SET domain of MLL1 (Supplementary Figure 18), a structural model of the EZH2 SET domain

was generated (Supplementary Figure 19a). Two of the three mutated residues within this domain are located near the putative S-Adenosylmethionine cofactor binding site (Asn 688 and Ser 690) of this structural model, while the third (Ile 646) is located near the putative substrate binding site, and the fourth (Arg 679) is located on a strand opposite the substrate binding pocket.

Isoleucine 646 is located within one strand of a three-strand anti-parallel β-sheet that forms the base of the lysine-containing peptide substrate binding pocket. The carbonyl and amide groups of Ile 646 participate in hydrogen bonds typical of an anti-parallel β-sheet and its aliphatic side chain interacts with several surrounding hydrophobic residues (Val 674 and Gln 648) (Supplementary Figure 19b). Mutation of the naturally occurring Ile residue to a Phe, as observed in several T-ALL patients, appears to create only minor steric clashes. The Ile residue is fully conserved in EZH2, MLL1 and other SET domains[49]. Substitution of larger amino acids at this position might subtly alter the stability of the protein, potentially affecting N-methyltransferase function.

The side chain of Arg 679 is located in a position that is remote with respect to the protein cofactor and substrate binding pockets (Supplementary Figure 19c). This Arg residue is found within a structured loop but the identity and number of amino acids in this loop varies between SET domains[49]. The consequences of this mutation are not immediately apparent. However, this model corresponds to only one of several domains of the full-length EZH2 protein which further functions in the context of the multi-protein polycomb repressor complex 2; therefore, mutation of Arg 679 to His may affect intra- or inter-protein domain-domain interactions not represented in our structural model.

The structural model suggests that Asn 688 is located in a loop structure that is predicted to bind to the edge of the adenine moiety of the S-Adenosylmethionine cofactor required for catalysis[49]. Mutation of Arg 685, adjacent to Asn 688 within the cofactor binding site, to a His or a Cys amino acid has been observed in patients diagnosed with myeloid malignancies and has been demonstrated to impair tri-N-methyltransferase function toward histone H3 at lysine 27[50,51]. These authors did not speculate on the mechanism of loss of enzymatic function but our model suggests that mutation of Arg 685 in the EZH2 SET domain would be structurally deleterious through perturbation of the cofactor binding site. Backbone and side chain atoms of Asn 688 interact with several residues within this region of the SET domain fold. The side chain δ-N of Asn 688 contacts the side chain of Glu 721 and backbone carbonyls of Ala 622 and Trp 624 (Supplementary Figure 19d). Further, the δ-O of the side chain of Asn688 has the potential to interact with the backbone amide group of the methionine moiety

within the bound S-Adenosylmethionine cofactor (modelled as the S-Adenosylhomocysteine reaction product here). Mutation of Asn 688 to Tyr, as observed in T-ALL patients in this study, would eliminate the putative stabilizing interactions noted above as well as create steric clashes due to the larger size of the Tyr *versus* Asn side chain. These structural perturbations would likely significantly destabilize this region of the EZH2 SET domain, potentially influencing cofactor binding and N-methyltransferase function.

Similar to Ile 646 and Asn 688, Ser 690 is fully conserved in EZH2 and other SET domains[49]. Ser 690 participates in a turn structure in a region of the SET domain that is proximal to the binding site for the amine group of the adenosine moiety of the S-Adenosylmethionine cofactor (modelled here as the S-Adenosylhomocysteine reaction product). The backbone amide nitrogen of Ser 690 interacts with the carbonyl of Phe 724, while its side chain hydroxyl group interacts with the backbone amide group of residue Asn 692 (Supplementary Figure 19e). Mutation of this Ser residue to a Leu could result in the loss of stability within the loop at the base of the binding site for the adenine moiety of the S-Adenosylmethionine cofactor. This may destabilize this region of the SET domain structure and potentially influence cofactor binding and N-methyltransferase function.

Mutations to the SET domain of EZH2 have previously been identified in a population of patients exhibiting follicular and diffuse large B-cell lymphomas[52]. One such mutation was observed in residue Tyr 641, which is conserved as an identity in many human SET domains, and has been demonstrated to be critical for methylation of Lys 27 of Histone H3 (ref. [52]). Morin, *et al.*, used methods essentially identical to those described herein to develop a homology-based structural model of the EZH2 SET domain bound to S-Adenosylhomocysteine and a histone-derived, lysine-containing peptide substrate. In this model, Tyr 641 was positioned in close proximity to the Lys side chain of the bound substrate peptide. This structural observation was consistent with biochemical results that showed loss of histone N-methyltransferase function (with a non-methylated lysine peptide substrate) upon substitution of this residue with either His, Asn, Ser or Phe (Supplementary Figure 19f).[52] Based on modelling, each of these substitutions was envisioned to alter substrate binding and N-methyltransferase function. Further investigation by Sneeringer *et al.,* showed that these mutations were associated with significantly decreased activity toward substrate peptides with non-methylated and mono-methylated lysine residues but that activity was increased toward a corresponding di-methylated lysine substrate[53]. These recent results, together with the modelling results, suggest that the noted mutations enlarge the substrate binding pocket, causing the non- and mono-methylated lysine side chains to be poorly accommodated by the N-methyltransferase active site. In

contrast, the side chain of a di-methylated lysine substrate may be well accommodated by the enlarged binding pocket, giving rise to enhanced catalytic activity.

In contrast to these mutations to Tyr 641 at the base of the substrate binding pocket, model-based structural analysis of the novel mutations in the EZH2 SET domain identified in this study in T-ALL patients suggests that they alter cofactor binding and/or cause local protein instability, both of which we predict would be associated with inhibition of N-methytransferase activity. For example, we speculate that one of the T-ALL-associated mutations (N688Y) would significantly perturb and destabilize the protein structure near the S-Adenosylmethionine cofactor binding site, adversely affecting cofactor binding and catalytic activity. Structural modelling also indicates that two other T-ALL-associated mutations (I646F and S690L) are likely to cause structural perturbations more subtle than those associated with the N688Y mutation. Nonetheless, we have identified potential mechanisms through which these mutations could destabilize the structure of localized regions of the SET domain, which, in turn, could alter overall domain stability and/or N-methyltransferase function. In summary, three of the T-ALL-associated mutations within the gene for the EZH2 protein occur within the SET N-methyltransferase domain and, to differing extents, are likely to negatively impact lysine N-methyltransferase activity.

SUPPLEMENTARY TABLES

Supplementary Table 1. Details of whole genome sequenced cases.

The 12 patients sequenced by WGS were diagnosed with ETP ALL based on immunophenotypic criteria.[54]

Sample	Sex	Race	Age at diagnosis	WCC	Karyotype	Outcome
SJTALL001	M	B	10	2.2	47,XY,der(1)add(1)(p36.3)add(1)(q32),t(4;7)(q21;p22),del(9)(p13),der(11)add(11)(p13)add(11)(q21),add(12)(p11.2),add(18)(q23),mar[12]/ 47,idem,-der(1),-der(11),add(19)(p13.3)[7]/ 46,XY[6]	Relapse with lineage switch at 13 months, died at 25 months post diagnosis
SJTALL002	F	W	6	63.8	47,X,t(X;12;2)(q13.1;12p12.2;2q33.3),del(9)(q13q32),+19[18]	BMT 9 months post diagnosis, died
SJTALL003	M	B	7	44.8	46,XY,del(1)(p22p32),inv(2)(p11.2q13)c,der(4)t(4;10)(q21;q26),der(5)t(5;10)(p13;p11.2),der(10)t(4;10;4;10;5)(4qter->4q?::10p11.2->10q11.2::4q?::10q?::5p13>5pter)[18]/46,XY,inv(2)(p11.2q13)c [2]	Haematologic relapse 8 months post diagnosis, died
SJTALL004	M	W	3	21.5	47,XY,+9,der(11)del(11)(p11.2p15)inv(11)(p11.2q22)[19]/46,XY[1]	Induction failure, BMT 5 months post diagnosis, alive
SJTALL005	F	B	12	16.0	46,XX,t(4;11)(q21;p15),del(5)(q22q35),del(12)(p12)[20]	BMT 4 months post diagnosis, alive
SJTALL006	M	W	10	52.3	46,XY	Alive
SJTALL007	M	W	18	13.9	46,XY,+6,inv(10)(p13q22),add(12)(p13),der(12)t(12;14)(p13;q11.2),del(13)(q22q24),-14[17]/46,XY [3]	Relapse 3 years post diagnosis, died
SJTALL008	M	W	2	3.0	46,XY	Cord blood transplant 11 months post diagnosis, relapse and BMT 3 years post diagnosis
SJTALL009	M	B	12	181.6	46,XY,add(7)(q35),del(11)(q21q23.1),del(13)(q12q22)[20].nuc ish[MLLx2]	Alive
SJTALL011	F	W	8	172.4	46,XX,t(1;6)(p32;q26),t(3;4)(p13;p16),del(6)(q13q21)[14]	Alive
SJTALL012	F	W	15	6.1	46,XX	BMT 4 months post diagnosis, alive
SJTALL013	F	B	13	4.0	47,XX,+4,t(10;12)(p11.2;q15)[3]/47,idem,del(5)(q22q35)[3]/46,XX[14]	Multiple relapses, BMT 6 years post diagnosis, died

Supplementary Table 2. DNA copy number alterations in ETP and non-ETP T-ALL identified by SNP array analysis

See table: “Table_S2_SJTALL_DNA_CNA_segments_annotated.xlsx”

Detailed listing of all segments of DNA copy number gains and losses identified by SNP array analysis for the entire ETP and non-ETP ALL cohort. Segments were identified by reference normalisation and circular binary segmentation (CBS) as previously described[38,55-58]. Segments obtained from CBS analysis of Affymetrix 250k Nsp and 250k Sty arrays (based on hg17 genome build) have been lifted over to hg18 (used for SNP 6.0 arrays).

Supplementary Table 3. Cumulative DNA copy number alteration and lesion number in ETP and non-ETP T-ALL identified by SNP array analysis

See table: “Table_S3_SJTALL_N_cumulative_lesion_summary.xlsx”

Supplementary Table 4. Results of DNA copy number alteration for ETP and non-ETP ALL using GISTIC.

See file “Table_S4_SJTALL_combined_GISTIC_results.xlsx”.

This file has 6 tabs, three for ETP and three for non-ETP ALL.

“(ETP/non-ETP) GISTIC scores”: gives all the GISTIC scores, -log10(q-value), average amplitudes among aberrant samples, and frequency of aberration, across the genome for both amplifications and deletions.

“(ETP/non-ETP) amplified genes”: gives detailed info for each significant region (those peaks above the green line in the Supplementary Figure 3), i.e. cytoband, q-value, interval, and refGene & microRNAs within the interval.

“(ETP/non-ETP) all lesions summarises all the significantly and recurrently amplified or deleted regions and for each region, listed segment mean info for each case”

-

Supplementary Table 5. Primer sequences for RT-PCR of chimeric fusion transcripts and rearrangements in ETP ALL.

Primer	Sequence (5' to 3')	Purpose
ETV6 F1 C2036	gacgggctgcatagggaagggaag	RT-PCR of ETV6-INO80D
INO80D R1 C2037	gcacaccatctgactgggcaaggac	RT-PCR of ETV6-INO80D
ETV6 F2 C2038	cacacacagccggaggtcatactgc	RT-PCR of ETV6-INO80D
INO80D R2 C2039	cgcccattcaatagctcccctaggtc	RT-PCR of ETV6-INO80D
BG201338 F1 C2040	caagaccggaaaggcagcaatagcc	RT-PCR of BG201338-ETV6
ETV6 R1 C2041	gccagtccgttgggatccactatcc	RT-PCR of BG201338-ETV6
BG201338 F2 C2042	ctgttcatcccccaagaccggaaag	RT-PCR of BG201338-ETV6
ETV6 R2	cgcagggctctggacattttctcat	RT-PCR of BG201338-ETV6
INO80D F1 C2044	agtgctgatgagttgccggatgaca	RT-PCR of INO80D-BG201338
BG201338 R1 C2045	ggagctccaaaggtggcagctgttc	RT-PCR of INO80D-BG201338
INO80D F2 C2046	caaaaacccattccacctgcagtcc	RT-PCR of INO80D-BG201338
BG201338 R2 C2047	gaggaagagaggaagccctggctga	RT-PCR of INO80D-BG201338
ETV6 F3 C2048	tctgggttggggagaggaaaggaaa	RT-PCR for full length cloning of ETV6-INO80D
INO80D R3 C2049	ccaaactgtggtatccttgggggttct	RT-PCR for full length cloning of ETV6-INO80D
NUP214 F1 C2700	agtggggccaagacatttggtggat	RT-PCR of NUP214-SQSTM1
SQSTM1 R1 C2701	agctgcttggctgtgagctgctctt	RT-PCR of NUP214-SQSTM1
NUP214 F2 C2702	cgtgtttgggtctggaaacactggaa	RT-PCR of NUP214-SQSTM1
SQSTM1 R2 C2703	ctctggcgggagatgtgggtacaag	RT-PCR of NUP214-SQSTM1
SQSTM1 F1 C2704	caggaaactggagcccacgtcctc	RT-PCR of SQSTM1-NUP214
NUP214 R1 C2705	gtaaaggctggggctgaaccgaatg	RT-PCR of SQSTM1-NUP214
NAP1L1 F1 C2380	tccttgctgcagacttcgaaattggtc	RT-PCR of NAP1L1-MLLT10
MLLT10 R1 C2381	gggcccaacccccattatctgttct	RT-PCR of NAP1L1-MLLT10
NAP1L1 F2 C2382	ccgcctggctccccatactagtcg	RT-PCR of NAP1L1-MLLT10
MLLT10 R2 C2383	tgcaaaggcagccagatgaagtgct	RT-PCR of NAP1L1-MLLT10
NAP1L1 F3 C2384	cccctcctgaagttcctgagagtgga	RT-PCR of NAP1L1-MLLT10
MLLT10 R3 C2385	gcaccagtggctgctttgctttctc	RT-PCR of NAP1L1-MLLT10
MLLT10 F1 C2386	agacgagagaggctgggccgagaac	RT-PCR of MLLT10-NAP1L1
NAP1L1 R1 C2387	ccagggaagcagaaggtagaccagtgc	RT-PCR of MLLT10-NAP1L1
INO80DF1 C3024	cctctttcccaggagcccatccac	Genomic PCR of SJTALL208
ETV6R1 C3025	tgaccagcgatagcctcacaaatcg	Genomic PCR of SJTALL208

Supplementary Table 6. Cases studied by gene expression profiling.

ETP cases	Non-ETP cases
SJTALL001	SJTALL015
SJTALL002	SJTALL016
SJTALL005	SJTALL019
SJTALL006	SJTALL020
SJTALL009	SJTALL021
SJTALL010	SJTALL022
SJTALL011	SJTALL024
SJTALL161	SJTALL025
SJTALL162	SJTALL026
SJTALL163	SJTALL027
SJTALL164	SJTALL028
SJTALL165	SJTALL029
	SJTALL030
	SJTALL031
	SJTALL032
	SJTALL033
	SJTALL034
	SJTALL036
	SJTALL037
	SJTALL038
	SJTALL041
	SJTALL044
	SJTALL047
	SJTALL048
	SJTALL050
	SJTALL056
	SJTALL066
	SJTALL067
	SJTALL068
	SJTALL069
	SJTALL070
	SJTALL071
	SJTALL072
	SJTALL075
	SJTALL077
	SJTALL078
	SJTALL079
	SJTALL143
	SJTALL157
	SJTALL159

Supplementary Table 7. Coverage data of cases sequenced.

Genomic Coverage: The average coverage of all non-ambiguous bases in hg18. Exon Coverage: The average coverage at all exonic bases (including all noncoding RNAs annotated in RefSeq). % Genomic bases covered: The percentage of all non-ambiguous bases covered at least 10x. % exonic bases covered: The percentage of all bases in RefSeq annotated exons covered at least 10x. % coding bases covered: The percentage of all RefSeq protein coding bases covered at least 10x

Patient	Germline / Diagnosis	Lanes	Nucleotides Sequenced	Genome Coverage	Haploid Coverage	Exon Coverage	% Genomic bases covered	% Exonic bases covered	% Coding bases covered	% SNP Detection
SJTALL001	G	22	129,535,642,600	25.8	26.9	23.7	97	92	92	96
SJTALL001	D	29	165,008,067,600	38.3	39.7	33.3	99	95	95	93
SJTALL002	G	16	96,274,939,400	24.3	26.1	21.2	97	89	88	98
SJTALL002	D	23	149,081,350,400	33.8	35.7	29.3	98	91	91	99
SJTALL003	G	16	117,372,461,000	25	26.5	21.2	96	84	82	99
SJTALL003	D	24	139,314,317,600	32.4	34.2	27.2	97	88	86	99
SJTALL004	G	18	108,334,958,400	26	27.5	22.2	96	86	84	99
SJTALL004	D	20	128,099,872,000	32.3	34.2	26.8	98	88	86	98
SJTALL005	G	16	107,795,802,500	26.5	28.6	21.6	96	81	78	98
SJTALL005	D	22	132,553,231,400	32.2	33.8	26.8	98	89	88	99
SJTALL006	G	16	107,335,770,600	25.9	27.7	22	96	85	83	99
SJTALL006	D	20	118,231,690,400	29.7	31.1	25.1	97	88	87	99
SJTALL007	G	16	111,923,423,200	28.4	30.1	23.8	96	83	81	99
SJTALL007	D	20	141,439,535,400	35.7	39.2	28.2	96	80	76	99
SJTALL008	G	16	112,490,929,000	27.8	30.6	23.1	94	79	75	99
SJTALL008	D	20	126,150,598,600	31.6	33.6	26	97	84	82	99
SJTALL009	G	20	108,240,770,200	25.7	27.7	21.8	96	85	84	98
SJTALL009	D	20	136,180,034,600	32.1	34.9	25.7	96	82	79	98
SJTALL011	G	16	102,342,708,400	23.9	24.9	20.4	97	86	84	99
SJTALL011	D	20	138,572,715,200	33.7	36.5	27	96	79	75	98
SJTALL012	G	16	108,380,991,000	28.3	30.5	23.5	96	79	76	99
SJTALL012	D	20	139,822,047,200	35.4	37.1	28.5	97	86	83	99
SJTALL013	G	19	113,603,985,200	28.6	29.3	25	98	89	87	99
SJTALL013	D	20	131,653,175,400	34.4	36.1	27.9	98	87	85	98

Supplementary Table 8. Coverage data for exome sequencing

Sample	>=10x	>=20x	>=30x
SJTALL169_D	94.2	90.4	87.0
SJTALL169_G	90.5	84.8	78.8
SJTALL192_D	94.4	90.4	86.7
SJTALL192_G	94.3	90.2	86.4
SJTALL208_D	92.2	87.1	81.9
SJTALL208_G	94.4	90.1	85.9

Supplementary Table 9. Summary of somatic sequence mutations, somatic structural variations and somatic copy number alterations identified in the 12 ETP ALL cases.

*All tier 1 mutations were validated. ^tiers2-4 only include mutations found by both WU and SJCRH methods. The average validation rate for these is 98%. However, the false negative rate is 23%. Tier 1: Coding synonymous, nonsynonymous, splice site, and non-coding RNA variants; 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); Tier 3: Variants in non-repeat masked regions; and Tier 4: The remaining SNVs. Combined SNV/SV refers to the union (non-redundant count) of SV and CNA. The number of structural variants and DNA copy number variants in each case are not the same as (1) some SVs are copy-neutral rearrangements; (2) CNVs from aneuploidy or gross alterations extending to the telomere will not be detected by WGS SV identification algorithms such as CREST; and (3) some CNV breakpoints are in repetitive or unmappable regions and will not be detected by SV algorithms.

Sample	tier1*	^tier2	^tier3	^tier4	Total	SVs	CNV (N) Amp	CNV (Mb) Amp	CNV (N) Del	CNV (Mb) Del	Combined CNV/SV (N)
SJTALL001	17	106	567	797	1,487	22	14	254	12	195	39
SJTALL002	15	81	433	521	1,050	24	16	122	21	146	42
SJTALL003	26	177	712	887	1,802	25	2	0.1	26	138	35
SJTALL004	10	66	344	415	835	5	1	140	6	39	10
SJTALL005	13	86	428	536	1,063	6	0	0	6	50	9
SJTALL006	9	53	230	275	567	0	0	0	1	11	1
SJTALL007	21	178	826	904	1,929	10	2	170	15	33	20
SJTALL008	6	29	88	112	235	0	0	0	0	0	0
SJTALL009	14	126	691	804	1,635	16	2	0.6	17	119	26
SJTALL011	12	49	283	367	711	15	1	44	5	46	20
SJTALL012	16	82	487	617	1,202	14	1	0.2	5	2	15
SJTALL013	16	106	468	576	1,166	9	1	191	2	71	11
All	175	1,139	5,557	6,811	13,682	146	40	922	116	849	228
mean	15	95	463	568	1,140	12	3	77	10	71	19
min	6	29	88	112	235	0	0	0	0	0	0
max	26	178	826	904	1,929	25	16	254	26	195	42

Supplementary Table 10. Comparison of mutation frequencies to previously published whole genome sequencing studies in cancer.

Disease	Region	N cases	Somatic Mutations in Coding Regions				Reference
			Non-silent		Silent		
			Total	# per case	Total	# per case	
Prostate	WGS	7	162	23	N/A	N/A	Berger et al, 2011[3]
AML	WGS	1	8	8	N/A	N/A	Ley et al, 2008[59]
AML	WGS	1	8	8	2	2	Mardis et al, 2009[60]
Breast Cancer	WGS	1	29	29	11	11	Ding et al, 2010[61]
Melanoma Cell line	WGS	1	187	187	102	102	Pleasance et al, 2010[62]
Lung cancer cell line	WGS	1	98	98	36	36	Pleasance et al, 2010[63]
Melanoma	RNA-seq	1	27	27	N/A	N/A	Berger et al, 2010[64]
Lung cancer	WGS	1	302	302	90	90	Lee et al, 2010[1]
Myeloma	WGS	23	793	34	N/A	N/A	Chapman et al, 2011[65]
	Exome	16	448	28			
ETP ALL	WGS	12	110	9	39	3	This study

Supplementary Table 11. Validated sequence mutations (including both substitution variants and indels) in RefSeq genes for WGS cases.

See Excel Table "Table_S11_validated_WGS_sequence_mutations.xls":

Transcriptome sequencing was carried out for SJTALL002 and SJTALL012. For these two samples, mutant alleles that are expressed (as determined by transcriptome sequencing) are highlighted in orange while those that were not expressed are highlighted in grey. Assessment of functional impact of missense mutations calculated by POLYPHEN and SIFT is included.

Exome sequencing was performed for SJTALL169, SJTALL192 and SJTALL208. Variations for these cases are highlighted in purple.

Column definition is listed below:
A: GeneName: HUGO gene symbolB: VarType: Sub=substitution, Indel=insertion/deletion
C: Sample: name of the sample
D: Chr: chromosome
E: HG18_Pos: chromosome position in hg18 coordinates.
F: Class: classification based on amino acid change pattern. Exon refers to mutations in non-coding RNA genes. splice_region refers to mutations not directly affect the canonical splice sites but are located within 10bp of the canonical splice sites .
G: AAChange: predicted amino acid change for the mutation
H: ProteinGI: NCBI protein GI number
I: mRNA_acc: Refseq accession number
J: #Mutant_In_Tumour: number of reads containing the mutant allele in WGS in tumour
K: #Total_In_Tumour: number of reads covering the site in WGS sequencing in tumour
L: #Mutant_In_Normal: number of reads containing the mutant allele in WGS in normal
M: #Total_In_Normal: number of reads covering the site in WGS in normal
N: Empty
O: #Mutant_In_Tumour_Validation: number of reads containing the mutant allele in 454 validation sequencing in tumour. A site validated by Sanger sequencing is marked "Sanger"
P: #Total_In_Tumour_Validation: number of reads covering the site in 454 validation sequencing in tumour. A site validated by Sanger sequencing is marked "Sanger"
Q: #Mutant_In_Normal_Validation: number of reads containing the mutant allele in 454 validation sequencing in normal. A site validated by Sanger sequencing is marked "Sanger"
R: #Total_In_Normal_Validation: number of reads covering the site in 454 validation sequencing in normal. A site validated by Sanger sequencing is marked "Sanger"
S: ReferenceAllele: the allele represented in the reference human genome. Reference allele is marked as – for an insertion
T: MutantAllele: mutant allele.
U: Flanking: 20bp[reference allele/mutant allele]20bp
V: SIFTResult: deleterious status assigned by SIFT
W: SIFTScore: SIFT score
X: pph2result: deleterious status assigned by polyPHEN2
Y: pph2score: PolyPHEN2 score

Supplementary Table 12. Numbers of sequence mutations in Refseq genes for WGS cases.

*Other mutations include those in non-coding RNAs and those within 10bp of the splice site but do not change the canonical splice site.

Case	Missense	Nonsense	Indel	Total Coding	Splice	Other*	Total non-silent	Silent	Total
SJTALL001	7	3	0	10	0	3	13	4	17
SJTALL002	9	0	1	10	2	1	13	3	16
SJTALL003	16	1	1	18	0	2	20	8	28
SJTALL004	7	0	0	7	0	2	9	1	10
SJTALL005	5	0	4	9	2	2	13	3	16
SJTALL006	6	1	1	8	0	0	8	2	10
SJTALL007	12	0	1	13	1	1	15	5	20
SJTALL008	3	0	0	3	0	1	4	2	6
SJTALL009	7	2	1	10	0	2	12	3	15
SJTALL011	9	0	2	11	0	1	12	2	14
SJTALL012	11	0	3	14	0	2	16	3	19
SJTALL013	10	1	4	15	0	3	18	3	21
Total (all cases)	102	8	18	128	5	20	153	39	192
Mean	8.5	0.67	1.5	10.67	0.42	1.67	12.76	3.25	16.01
Min	3	0	0	3	0	0	4	1	6
Max	16	3	4	18	2	3	20	8	28

Supplementary Table 13. **Validated structural variations found in the 12 WGS cases.**

See excel workbook of structural rearrangements:
"Table_S13_SJTALL_Structural_Variations.xls"

The column title and definition are listed below. Predicted Fusion gene (column J) was predicted based on the genomic coordinates of the breakpoints. Row 131 was highlighted in grey because the real gene fusion event cannot be predicted by using the genomic coordinates alone. The chr7/chr21 translocation generates a *RUNX1-EVX1* fusion based on RNA-seq data.

The table includes structural variants identified from analysis of exome sequencing data in cases SJTALL208 and SJTALL192

Column Title	Column Definition
Sample	Sample name
ChrA	Chromosome for breakpoint A
PosA	Position of breakpoint A
OrientationA	+ Region to the left of PosA is included in mutant genotype - Region to the right of PosA is included in mutant genotype
ChrB	Chromosome for breakpoint B
PosB	Position of breakpoint B
OrientationB	+ Region to the right of PosA is included in mutant genotype - Region to the left of PosA is included in mutant genotype
Type	INS Insertion DEL Deletion INV Inversion ITX Intrachromosomal translocation CTX Interchromosomal translocation
Usage	GENIC: Both endpoints were in genes: checked for fusion HALF_INTERGENIC: One endpoint was in a gene: checked for truncation CO_GENIC: Both endpoints were in genes: checked for and found fusion that involved multiple events INTERGENIC / INTRONIC: Neither endpoint was in a gene or both were in the same intron of a gene; no gene fusion or truncation INVERTED_REPEAT: Both endpoints were in the same gene, but in opposite orientations: checked for truncation
Gene	Fusion or truncated gene that would result from structural variation
Chromosomes	Chromosomes involved in the rearrangement
Tx	Number of predicted fusion transcripts
Valid CDS	Number of predicted fusion transcripts with an annotated CDS start and stop
In-Frame CDS	Number of "Valid CDS" transcripts with a CDS length divisible by three.
Modified In-Frame CDS	Number of "In-Frame CDS" transcripts that are not identical to an existing annotated transcript.

Column Title	Column Definition	
Non-template insertion	Non-reference bases inserted between the breakpoints	
Microhomology	Microhomology between the two breakpoints	
mutA	Number of reads supporting the structural variation at breakpoint A	
mutB	Number of reads supporting the structural variation at breakpoint B	
refA	Number of reads supporting the reference allele at breakpoint A	
refB	Number of reads supporting the reference allele at breakpoint B	
mut freq a	Mutant frequency at breakpoint A, calculated as mutA/(mutA + refA)	
mut freq b	Mutant frequency at breakpoint B, calculated as mutB/(mutB + refB)	
Possible mechanism	RSS NHEJ MHMT	RAG-mediated recombination signal sequences Non-homologous end join micro-homology mediated translocation

Reads were counted as reference-supporting reads if they met all of the following criteria:

- They had a continuously aligning block that aligned to at least one base on either side of the breakpoint.
- They were mapped to the same strand as the mutant-supporting reads. This requirement was included because all mutant-supporting reads would be on the same strand due to the behaviour of the BWA aligner. The BWA aligner will introduce soft-clipping only when performing Smith-Waterman realignment of unmapped reads in the neighbourhood of their already-mapped mates.
- Within the locally aligned block, the base mismatch rate on each side of the breakpoint was less than 50%. This requirement was included because some reads that actually supported the mutant allele would align a few bases across the breakpoint with large numbers of base mismatches on one side.

Supplementary Table 14. **Copy number alterations identified by analysis of WGS and SNP microarray data for the 12 discovery WGS ETP ALL cases.**

See table: “S14_Consolidated_WGSS_SNP_CNV”

Notes regarding column headers:

Class: Loss or Gain prediction

Estimated change from RegTree: Estimated copy number change (CN in D – CN in G) using CONSERTING

Tier: 1 if CNV is called both in NGS by CONSERTING and by SNP array; 2 if it is called by only one approach/algorithm

Calling Algorithm: the calling algorithm (regTree: CONSERTING in NGS, SNPArray: CBS in SNP array) that predicts the CNV for this tier 2 region.

Supplementary Table 15. **Correlation between abnormalities detected by cytogenetics and DNA copy number alteration / structural variant analysis of WGS data.**

See table: “Table_S15_cytogenetics_correlation.xlsx”

Supplementary Table 16. **Chimaeric fusion genes predicted by CREST analysis of whole genome sequencing and RNA-seq data.**

**CTNNA3* is not expressed in this sample on microarray-based gene expression profiling. ***BG201338* is a non-Refseq EST without a known open reading frame. This fusion is predicted to encode a chimeric protein.

Sample	Fusion Gene	Validated	In frame chimaeric fusion?
SJTALL002	*ETV6-INO80D*	RT-PCR	Yes
SJTALL002	*BG201338-ETV6*	RT-PCR	No
SJTALL002	*INO80D-BG201338*	RT-PCR	Yes**
SJTALL003	*CTNNA3-ARHGAP21*	No (RT-PCR)*	
SJTALL009	*SQSTM1-NUP214*	RT-PCR	Yes
SJTALL009	*NUP214-SQSTM1*	RT-PCR	Yes
SJTALL012	*RUNX1-EVX1*	RNA-seq	Yes
SJTALL012	*NDST2-RUNX1*	No (RT-PCR)	
SJTALL013	*NAP1L1-MLLT10*	RT-PCR	Yes
SJTALL013	*MLLT10-NAP1L1*	RT-PCR	Yes

Supplementary Table 17. **List of all non-silent validated mutations from WGS (n=12) and recurrence screening (n=94).**

See excel table: “Table_S17_validated_mutations_recurrence_genes.xls”.

Supplementary Table 18. Matrix showing copy number alterations and sequence mutations for all cases.

See Excel table: “Table_S18_Matrix_CNA_mutation_all_cases.xlsx”

Supplementary Table 19. Frequency of recurring somatic genetic alterations in ETP and non-ETP T-ALL.

Data are shown for all cases examined, including 12 WGS cases, and 94 cases studied by Sanger sequencing of 42 candidate genes and SNP microarray data. Genes are ranked by their frequency of occurrence in ETP ALL. Histone modification genes sequenced include *EED, EZH2, SUZ12, SETD2* and *EP300.*

Gene	ETP (N,%) N=64				Non-ETP (N,%) N=42				*P*
	Sequence mutation only	Deletion only	Mutation and deletion	Total N (%)	Sequence mutation only	Deletion only	Mutation and deletion	Total N (%)	
ETV6	8	13	0	21 (33)	1	3	0	4 (10)	0.0050
WT1	16	2	0	18 (28)	3	1	1	5 (12)	0.0563
CDKN2A	0	16	0	16 (25)	0	34	0	34 (81)	<0.0001
PHF6	12	4	0	16 (25)	9	0	0	9 (21)	0.8159
DNM2	13	0	0	13 (20)	4	0	0	4 (10)	0.1802
NRAS	11	0	0	11 (17)	4	0	0	4 (10)	0.3943
SUZ12	2	9	0	11 (17)	0	1	0	1 (2)	0.0030
EZH2	6	1	3	10 (16)	0	2	0	2 (12)	0.5587
NOTCH1	10	0	0	10 (16)	18	0	0	18 (43)	0.0031
RUNX1	6	2	2	10 (16)	1	1	0	2 (5)	0.1172
FLT3	9	0	0	9 (14)	0	0	0	0	0.0108
EED	2	5	1	8 (13)	0	3	0	3 (7)	0.0208
IKZF1	2	6	0	8 (13)	0	1	0	1 (2)	0.0840
JAK3	7	0	0	7 (11)	0	0	0	0	0.0404
NF1	0	6	1	7 (11)	1	1	0	2 (5)	0.3134
GATA3	4	0	2	6 (9)	0	0	0	0	0.0797
PTEN	2	2	2	6 (9)	5	4	2	11 (26)	0.0299
CTCF	1	4	0	5 (8)	0	4	0	4 (10)	0.7379
ECT2L	5	0	0	5 (8)	3	0	0	3 (7)	1
IL7R	5	0	0	5 (8)	2	0	0	2 (5)	0.7007
JAK1	5	0	0	5 (8)	1	0	0	1 (2)	0.3989
SETD2	4	1	0	5 (8)	0	0	0	0	0.1543
RELN	4	0	0	4 (6)	2	0	0	2 (5)	1
SH2B3	3	1	0	4 (6)	0	0	0	0	0.1504
BCL11B	3	0	0	3 (5)	3	0	0	3 (7)	0.6793
EP300	3	0	0	3 (5)	0	0	0	0	0.2755
FBXW7	3	0	0	3 (5)	2	4	0	6 (14)	0.0856
PTPN11	3	0	0	3 (5)	0	0	0	0	
KRAS	2	0	0	2 (3)	0	0	0	0	0.52
BRAF	1	0	0	1 (2)	0	0	0	0	0.5170
DCLRE1C	1	0	0	1 (2)	0	0	0	0	-
HIST1H1B	1	0	0	1 (2)	0	0	0	0	-
HNRNPA1	1	0	0	1 (2)	0	0	0	0	-
HNRNPR	1	0	0	1 (2)	0	0	0	0	-
Development	-	-	-	37 (58)	-	-	-	7 (17)	<0.0001
Signalling	-	-	-	43 (67)	-	-	-	8 (19)	<0.0001
Histone modification	-	-	-	31 (48)	-	-	-	5 (12)	0.0001

Supplementary Table 20. Frequency of cytokine/Ras signalling and developmental mutations in ETP and non-ETP T-ALL

The frequency of alterations targeting lymphoid and haemopoietic development and cytokine receptor / Ras pathway signalling in ETP and non-ETP T-ALL cases, as defined by whole genome sequence data for 12 cases, and recurrence mutation screening and SNP array analysis for the recurrence cohort. *Genes considered were *EP300, ETV6*, *GATA2, GATA3, IKZF1* and *RUNX1*). **Genes considered were *NRAS, FLT3, JAK3, IL7R, JAK1, KRAS SH2B3, NF1, PTPN11,* and *BRAF*).

	ETP N=64 N (%)	Non-ETP N=42 N (%)	*P*
Development* (no signalling)	9 (14.1)	5 (11.9)	NS
Development (± signalling)	37 (57.8)	7 (16.7)	<0.0001
Signalling** (no development)	15 (23.4)	6 (14.3)	0.08
Signalling (± development)	43 (67.2)	8 (19)	<0.0001
Development and Signalling	28 (43.8)	2 (4.8)	<0.0001
Development or signalling	52 (81.3)	13 (31)	<0.0001
Neither	12 (18.8)	29 (69)	<0.0001

Supplementary Table 21. **Lack of association between the IL7R rs6897932 (T244I) genotype and the risk of T-ALL.**

Allele frequencies are shown for the recurrence cohort, and the recurrence cohort and 12 cases subjected to whole genome sequencing (WGS) combined. The genotype of the recurrence samples were based on analysis of Sanger sequencing data for the recurrence cohort. Using allelic chi-square test, there is no significant difference between the allelic frequency in the T-ALL samples and that of the HapMap CEPH population either using recurrence-only (*P*=0.6) or recurrence+WGS (*P*=0.51).

Population	C	Freq	T	Freq
Recurrence	136	0.73	50	0.27
Recurrence+WGS	153	0.73	57	0.27
HapMap-CEU	171	0.76	55	0.24
HapMap-JPT	141	0.82	42	0.18
HapMap-YRI	215	0.95	11	0.05
MS-US-control		0.74		0.27
MS-US		0.78		0.22
MS-EU-control		0.72		0.28
MS-EU		0.76		0.24

Supplementary Table 22. Sequence variants identified on analysis of WGS data obtained from matched non-tumour samples

See workbook: “Table_S22_sequence_variants_matched_normal_samples.xls”

The workbook contains 4 sheets.

Worksheet "Type", summarizes frequencies for different variant types

Worksheet "Gene" lists all genes with putative inherited variants in each sample, plus three columns (Total occurrences; isCancerGene: in Cancer Gene Census; hasSomatic: present as a somatic variant in sequencing of the recurrence cohort). Variants predicted to be deleterious by PolyPhen are underlined; those predicted to be deleterious by SIFT are marked in bold.

Worksheet "SomaticCancer", is a subset of tab "Gene" when the column "isCancerGene" or "hasSomatic" is "Yes". Here, the germline and somatic variants are separated by a forward slash (/) in the form of Germline AAChange / Somatic AAChange. If a sample has more than one event in a type, they are separated by commas.

Worksheet "Variants", lists the amino acid changes in each sample, the number of cases affected and the dbSNP ID. Note some variants present in dbSNP were retained because they were found in COSMIC (although the variants in COSMIC may not be somatic variants, but may be inherited). In each cell, the first letter is the reference allele, followed by two letters indicating if it's homozygous or heterozygous. Results of SIFT and PolyPhen analyses are also shown.

Supplementary Table 23. Listing of pathway analysis results for the WGS cases

See supplementary excel table: “Table_S23_SJTALL_WGS_pathway_GRIN_results.xlsx”

The meaning of the table column headers is as follows:

gset.source	source of gene-set definition
gset.id	gene-set identifier
gset.name	gene-set name
ngenes.gset	number of genes in the gene-set
gset.genes	list of genes in the gene-set
geneloc.not.found	list of gene-set genes without matched location data
nsubj.hit	number of subjects with at least one lesion overlapping a gene-set gene
pval.nsubj.hit	*P* value for the number of subjects with at least one lesion overlapping a gene-set gene
tot.nhits	total number of occurrences that a lesion overlaps a gene across the entire cohort
pval.tot.nhits	*P* value for tot.nhits
genes.hit	list of hit genes with number of subjects showing that hit
subjects.hit	list of subjects with number of hits observed
(SJTALL001).ghit	list of genes with number of hits for subject TALL001
(SJTALL001).nhit	number of hits for subject TALL001
(SJTALL001).pval	*P* value for number of hits for TALL001
(SJTALL001).sim.chr	list of chromosomes that Monte Carlo simulation was used because exact calculations were computationally infeasible

Supplementary Table 24. *limma* gene expression signature of 12 ETP versus 40 non-ETP T-ALL samples.

See supplementary excel table: “Table_S24_12SJETP_40SJnonETP_limma.xlsx”

Supplementary Table 25. Reverse engineering of transcription factor networks in ETP ALL using ARACNE.

See workbook: “Table_S25_ARACNE_analysis.xlsx”

Sheet “Limmanet”: ARACNE list of 30 regulators that control most differentially expressed genes (35,000 probe sets from *limma* analysis).

Sheet “Pathways”: DAVID (http://david.abcc.ncifcrf.gov) functional annotation clustering of 30 most significant regulators listed in the table 1.

Sheet “Regulons”. List of regulated genes for the 30 most significant regulators listed in the sheet “Network”.

Sheet “Correlation”. Spearman's rank correlation between “eigengene” generated based on a particular regulon (each regulon is regulated by a gene in the first column) and binary vector (1 – ETP; 0 – non-ETP).

Supplementary Table 26. Associations between ETP status, genetic lesions and outcome in T-ALL.

Estimates refer to cumulative incidence of relapse. *P* values from Gray's test.

Factors present/absent	n	Estimate + SE (%) Year 2	Year 5	p-value*
ETP				
No	41	2.4 (2.4)	8.2 (4.6)	<.0001
Yes	61	29.0 (6.3)	50.2 (8.4)	
BCL11B				
No	96	17.6 (4.0)	30.8 (5.4)	0.8885
Yes	6	16.7 (16.7)	16.7 (16.7)	
BIALLELIC *TRG*@ deletion				
No	32	21.3 (8.0)	44.7 (12.0)	0.1336
Yes	70	16.1 (4.5)	25.4 (5.7)	
***CDKN2A* deletion**				
No	55	21.7 (5.9)	35.8 (7.7)	0.2814
Yes	47	13.0 (5.0)	24.2 (7.0)	
DNM2				
No	85	16.1 (4.1)	29.2 (5.6)	0.5072
Yes	17	25.3 (11.5)	32.8 (12.7)	
ECT2L				
No	94	18.0 (4.1)	29.9 (5.4)	0.7670
Yes	8	12.5 (12.5)	34.4 (23.8)	
EED				
No	91	18.6 (4.2)	31.3 (5.6)	0.6171
Yes	11	10.0 (10.0)	20.0 (13.4)	
EP300				
No	99	18.2 (4.0)	29.4 (5.2)	0.8851
Yes	3	0	50.0 (50.0)	
ETV6				
No	78	18.6 (4.5)	27.7 (5.7)	0.3037
Yes	24	14.3 (7.9)	39.8 (12.7)	
EZH2				
No	90	14.1 (3.8)	26.6 (5.4)	0.0143
Yes	12	46.4 (16.8)	59.8 (18.4)	
FLT3				
No	93	19.2 (4.2)	32.0 (5.4)	0.0818
Yes	9	0	0	
GATA3				
No	97	16.3 (3.9)	27.8 (5.2)	0.0205
Yes	5	46.7 (29.8)	No Data	
IL7R				
No	95	17.9 (4.1)	29.7 (5.3)	0.8225
Yes	7	14.3 (14.3)	35.7 (24.0)	
JAK1				
No	97	15.3 (3.8)	28.6 (5.3)	0.1038
Yes	5	60.0 (26.0)	60.0 (26.0)	
JAK3				
No	95	16.6 (3.9)	28.2 (5.2)	0.0832
Yes	7	31.4 (20.6)	65.7 (35.8)	
LMO2 deletion				
No	97	18.6 (4.1)	32.2 (5.5)	0.7936
Yes	5	0	0	
MYB amplification				
No	96	17.7 (4.1)	30.8 (5.4)	0.9858
Yes	6	16.7 (16.7)	16.7 (16.7)	

Factors present/absent	n	Estimate + SE (%) Year 2	Year 5	p-value*
NF1				
No	94	18.0 (4.1)	29.0 (5.4)	0.2561
Yes	8	12.5 (12.5)	41.7 (20.5)	
NOTCH1				
No	74	18.8 (4.8)	35.9 (6.9)	0.2388
Yes	28	14.6 (6.9)	18.5 (7.6)	
FBXW7				
No	93	17.1 (4.1)	31.3 (5.6)	0.5841
Yes	9	22.2 (14.8)	22.2 (14.8)	
NOTCH_FBXW7				
NOTCH1 + FBXW7 +	5	40.0 (25.0)	40.0 (25.0)	0.1532
NOTCH1 + FBXW7 -	23	8.7 (6.0)	13.5 (7.5)	
NOTCH1 - FBXW7 +	4	0	0	
NOTCH1 - FBXW7 -	70	19.9 (5.0)	38.5 (7.3)	
NRAS				
No	87	18.0 (4.3)	28.2 (5.5)	0.4395
Yes	15	15.2 (10.4)	43.4 (16.3)	
PHF6				
No	77	17.8 (4.5)	32.5 (6.2)	0.6146
Yes	25	17.2 (8.1)	22.7 (9.3)	
PTEN mutation only				
No	91	17.3 (4.1)	31.6 (5.6)	0.9398
Yes	11	20.0 (13.4)	20.0 (13.4)	
PTEN deletion only				
No	93	15.9 (3.9)	26.2 (5.1)	0.0040
Yes	9	33.3 (16.8)	58.3 (19.3)	
PTEN any				
No	86	17.2 (4.2)	28.4 (5.4)	0.4046
Yes	16	20.0 (10.7)	38.9 (15.7)	
RB1				
No	95	19.0 (4.2)	31.4 (5.5)	0.2673
Yes	7	0	14.3 (14.3)	
RELN				
No	96	17.6 (4.0)	29.6 (5.4)	0.9190
Yes	6	16.7 (16.7)	37.5 (24.3)	
RUNX1				
No	90	16.4 (4.0)	30.0 (5.5)	0.9100
Yes	12	25.0 (13.1)	25.0 (13.1)	
SET-NUP214				
No	99	18.2 (4.0)	31.1 (5.3)	0.2292
Yes	3	0	0	
SH2B3				
No	98	15.2 (3.8)	28.1 (5.2)	0.0012
Yes	4	75.0 (29.2)	No Data	
***SUZ12* SV**				
No	93	17.0 (4.0)	27.9 (5.3)	0.1718
Yes	9	23.8 (15.9)	54.3 (21.3)	
***SUZ12* mutation**				
No	100	17.0 (3.9)	29.8 (5.3)	0.1929
Yes	2	50.0 (50.0)	50.0 (50.0)	
***SUZ12* SV or mutation**				
No	91	16.3 (4.0)	27.5 (5.4)	0.0676
Yes	11	28.4 (14.8)	52.3 (18.2)	
WT1				
No	79	15.8 (4.2)	29.6 (5.8)	0.9594
Yes	23	25.0 (10.1)	31.2 (11.1)	
Any development				

Factors present/absent	n	Estimate + SE (%) Year 2	Year 5	p-value*
No	60	12.1 (4.3)	23.3 (6.2)	0.0396
Yes	42	26.2 (7.3)	41.0 (9.1)	
Any signalling				
No	54	11.9 (4.6)	21.9 (6.4)	0.0446
Yes	48	24.2 (6.5)	40.4 (8.5)	
Polycomb repressor complex 2 / *SETD2*				
0 lesions	69	9.0 (3.5)	19.8 (5.7)	0.0007
1-4 lesions	33	36.3 (9.0)	52.1 (9.9)	

Supplementary Table 27. Prognostic effect of genetic alterations after adjusting for ETP status. CI, confidence interval. HR, hazard ratio.

	P	HR	HR 95% CI
ETP yes vs. no	0.00004	6.44	2.65-15.7
Haemopoietic/lymphoid development no vs. yes	0.22	0.61	0.29-1.34
1-9 years vs. older	0.032	2.33	1.08-5.04
ETP yes vs. no	0.00003	6.83	2.75-17.0
Signalling no vs. yes	0.79	0.90	0.43-1.90
1-9 years vs. older	0.064	1.95	0.96-3.96
ETP yes vs. no	0.00002	6.68	2.78-16.1
EZH2 alteration	0.086	2.19	0.90-5.35
1-9 years vs. older	0.034	2.22	1.06-4.67
ETP yes vs. no	0.00001	7.05	2.93-17.0
SUZ12 **alteration**	0.11	2.23	0.83-6.02
1-9 years vs. older	0.046	2.06	1.01-4.19
ETP yes vs. no	0.00002	6.68	2.78-16.1
SH2B3 **alteration**	0.011	5.01	1.44-17.5
1-9 years vs. older	0.068	1.91	0.95-3.85
ETP	0.0002	5.71	2.31-14.1
0 lesions in **PRC2** vs. 1-3 Lesions in RRC2	0.10	0.55	0.27-1.12
1-9 years	0.073	1.90	0.94-3.83

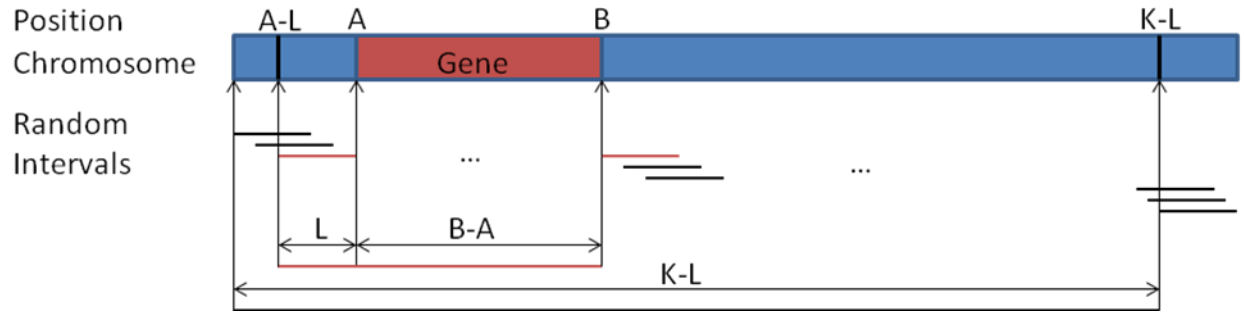
Supplementary Table 28. Comparison of the spectrum of genetic alterations in ALL and AML.

Patterns and frequency of genetic alterations in genes in this study are compared between (T-) ALL and AML. For more detailed reviews of the molecular genetics of AML, the reader is referred to other publications, such as Marcucci et al.[66]. CN-AML: cytogenetically normal AML.

	ALL	AML
BCL11B	Mutated and rearranged in T-ALL[67]	Uncommonly rearranged, no sequence mutations described.[68]
CEBPA	Not known to be mutated in ALL	Master regulatory transcription factor, mutated in 10-18% cytogenetically normal AML, commonly biallelic.[69]
EP300	Rarely mutated in relapsed B- and T-ALL. Mutated in lymphoma[25]	Sequence mutations not studied. Rarely target of translocation.[70]
ETV6	Common target of rearrangement (*ETV6-RUNX1* in B-progenitor ALL).[12] Focal deletions in B-progenitor ALL and in relapsed ALL.[71,72] Mutated in high risk B-progenitor ALL.[73]	Uncommonly rearranged in AML and myeloproliferative neoplasms. Sequence mutations in ~2% AML.[13,74]
FLT3	Mutated in <5% of ALL, but enriched in hyperdiploid, *MLL*-rearranged and high risk B-ALL[75,76]	Mutated (commonly ITD) in 28-34% AML
GATA3	This study is the first report of sequence mutations in deletions in T-ALL	No data
IKZF1	Deleted/mutated in ~15% B-progenitor ALL. Deletions/mutations very common BCR-ABL1 lymphoid leukaemia and novel subtype of BCR-ABL1-like ALL. Commonly mutated in mouse models of T-ALL.[33] Previously, mutations of *IKZF1* uncommon in human T-ALL samples.[38,77]	Not known to be mutated in AML; mutated in blast phase of JAK2-negative myeloproliferative neoplasms.[78]
IL7R	Mutated in T- and B-ALL[79]	No data
JAK1/2/3	*JAK1* mutations in T-ALL[80] and high-risk B-ALL.[36] JAK2 mutations in Down syndrome and high-risk B-ALL (especially at R683)[36,81]	Uncommonly mutated in AML.[82] JAK3 mutations in megakaryoblastic leukaemia.[83] JAK2 mutations in myeloproliferative neoplasms and at transformation to AML[84]
NRAS, KRAS	Mutated in ~20% B-progenitor ALL and up to 50% of *MLL*-rearranged and high hyperdiploid ALL.	Mutated in 9-14% CN-AML; higher frequency in AML with rearrangements of genes encoding the core-binding factor transcription complex (*RUNX1 (AML1)* and *RUNX1T1 (ETO))*
PTPN11	Mutated in 7% B-progenitor ALL but not T-ALL[85-88]	Mutated in juvenile myelomonocytic leukaemia[89] and uncommonly in AML[90]
RUNX1	Common target of rearrangement (*ETV6-RUNX1* in B-progenitor ALL).[12] Rare reports of T-ALL arising from familial platelet disorder.[91] Amplified in subset of B-ALL.[92] Not previously known to be mutated in ALL	Common target of rearrangement (e.g. *RUNX1-RUNX1T1*).[43] Somatic sequence mutations in 5-13% AML,[45,93] inherited mutations in FPD/AML.[47]
WT1	Commonly mutated in T-ALL	Mutations in 10-13% cytogenetically normal AML
EED	No prior data	No prior data
EZH2	No prior reports in ALL. Y641 mutations in follicular lymphoma	Non-Y641 inactivating mutations in myelodysplastic and myeloproliferative disorders[50,51]
SUZ12	No prior data	No prior data

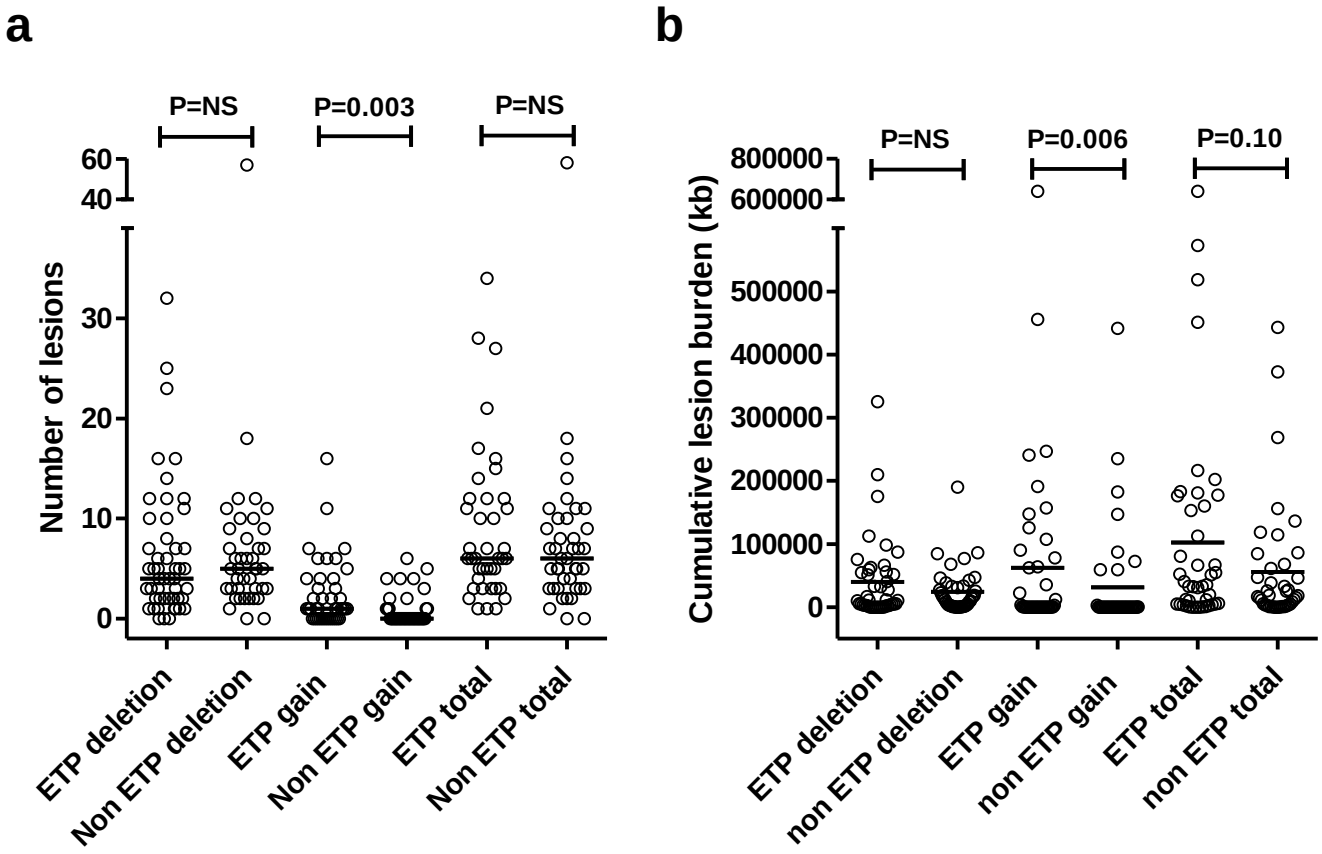
SUPPLEMENTARY FIGURES

Supplementary Figure 1. The genomic random interval (GRIN) model



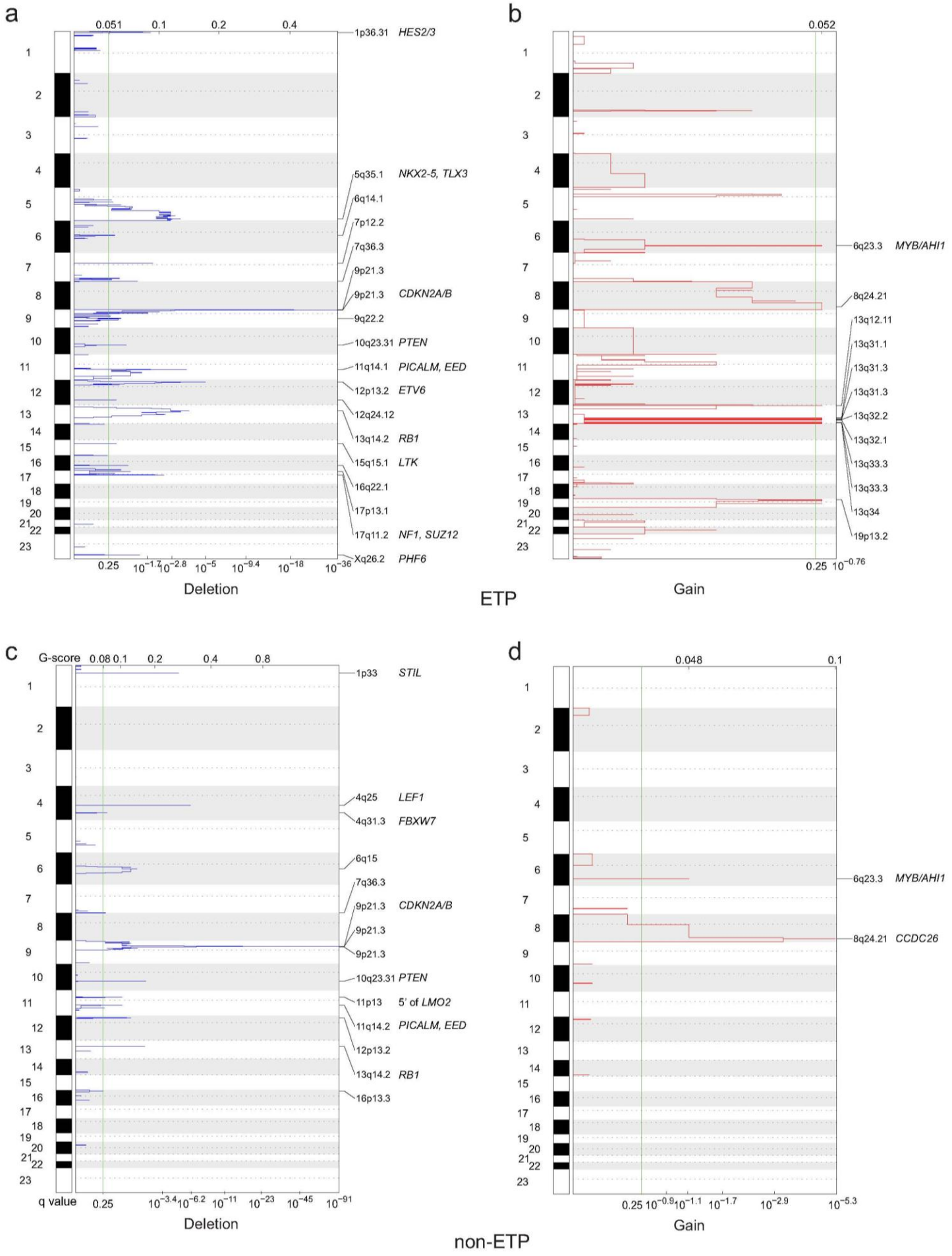
A lesion of length *L* on a chromosome of length *K* is modelled as having random start loci uniformly distributed between 1 and *K-L*. In the example above, the probability that the random interval of length *L* overlaps a gene at (*A*,*B*) is (*B-A+L*)/(*K-L*).

Supplementary Figure 2. Number and burden of DNA copy number alterations in ETP and non-ETP ALL.

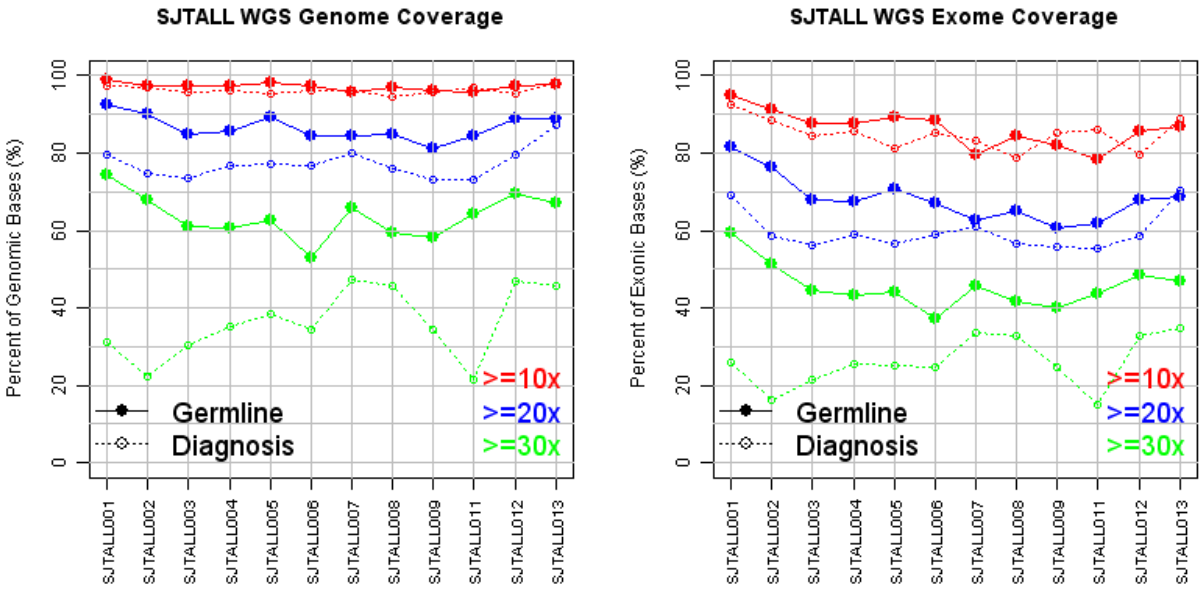


Supplementary Figure 3. Regions of significant DNA copy number alteration (GISTIC).

For each panel, the y-axis indicates markers for each chromosome along the genome, the top x-axis indicates GISTIC scores and the bottom x-axis indicates q-values (FDR). The green line represents default q-value cutoff of 0.25. The dotted lines within each chromosome separate p from q arms.


a
0.051
0.1
0.2
0.4
1p36.31 HES2/3
5q35.1 NKX2-5, TLX3
6q14.1
7p12.2
7q36.3
9p21.3
9p21.3 CDKN2A/B
9q22.2
10q23.31 PTEN
11q14.1 PICALM, EED
12p13.2 ETV6
12q24.12
13q14.2 RB1
15q15.1 LTK
16q22.1
17p13.1
17q11.2 NF1, SUZ12
Xq26.2 PHF6
0.25
Deletion
b
0.052
6q23.3 MYB/AHI1
8q24.21
13q12.11
13q31.1
13q31.3
13q31.3
13q32.2
13q32.1
13q33.3
13q33.3
13q34
19p13.2
0.25
Gain
ETP
c
G-score
0.08
0.1
0.2
0.4
0.8
1p33 STIL
4q25 LEF1
4q31.3 FBXW7
6q15
7q36.3
9p21.3 CDKN2A/B
9p21.3
9p21.3
10q23.31 PTEN
11p13 5' of LMO2
11q14.2 PICALM, EED
12p13.2
13q14.2 RB1
16p13.3
q value
0.25
Deletion
d
0.048
0.1
6q23.3 MYB/AHI1
8q24.21 CCDC26
0.25
Gain
non-ETP

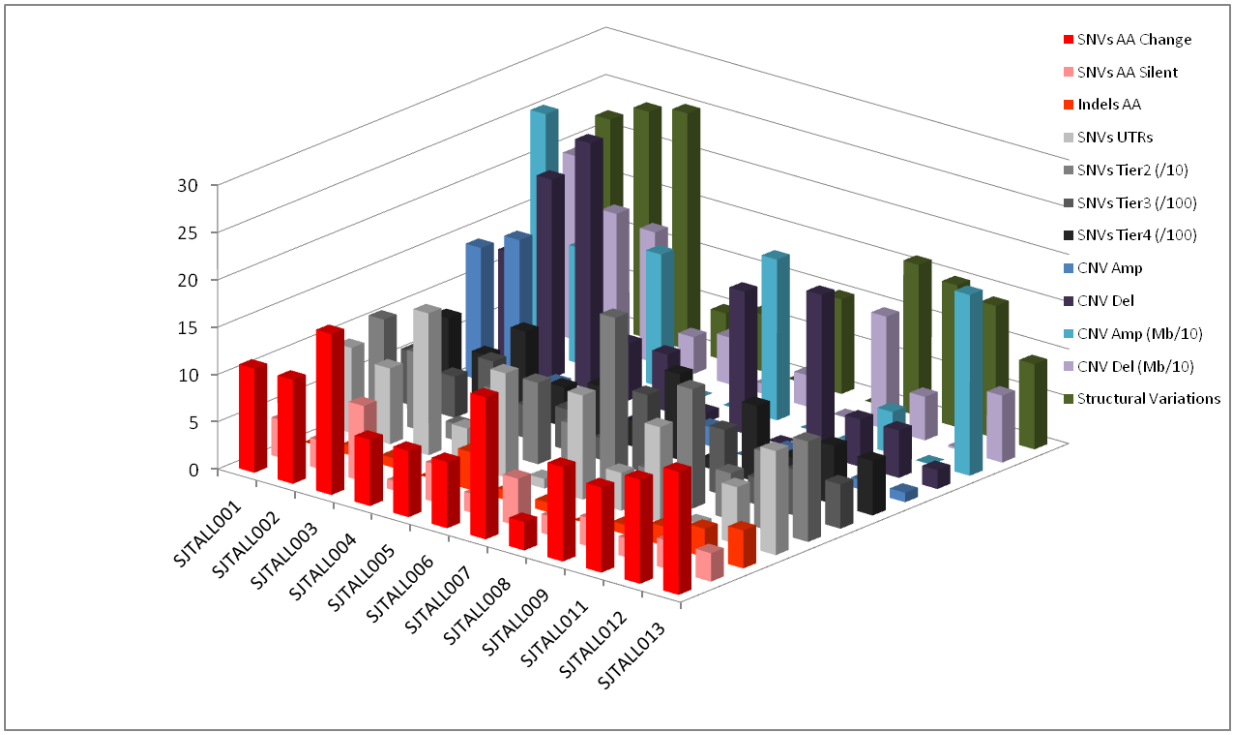
Supplementary Figure 4. Genome coverage of WGS cases



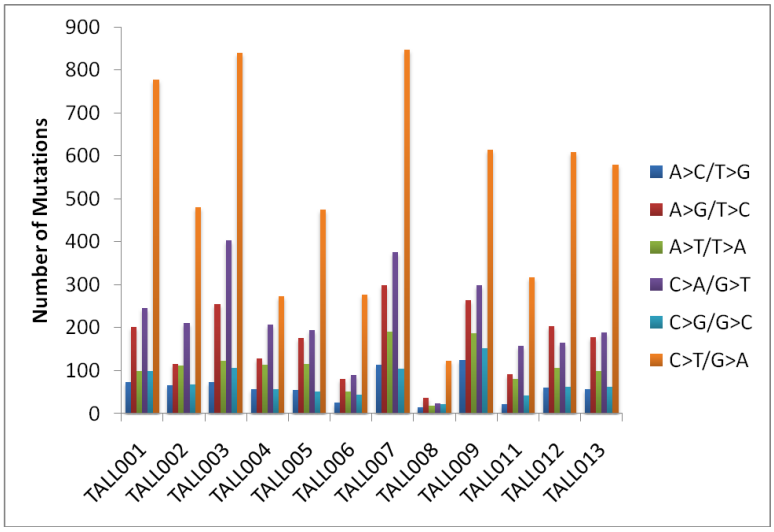
Supplementary Figure 5. The mutation spectrum of ETP ALL.

a, Numbers of sequence and structural variants. **b**, The raw number of sequence mutation counts for each ETP ALL case. **c**, The mutation frequency normalised (percentage) to the total for each sample. **d**, The fraction of mutations in the CpGs to NpGs (where N is ATCG). The background fraction in *c* is obtained from ref. [63]. The mutation spectrum is similar to that observed in acute myeloid leukaemia genomes[59].

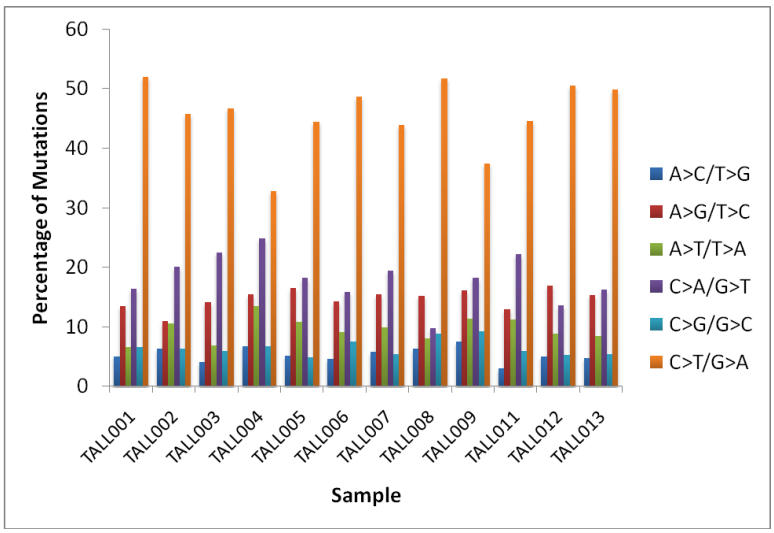
a



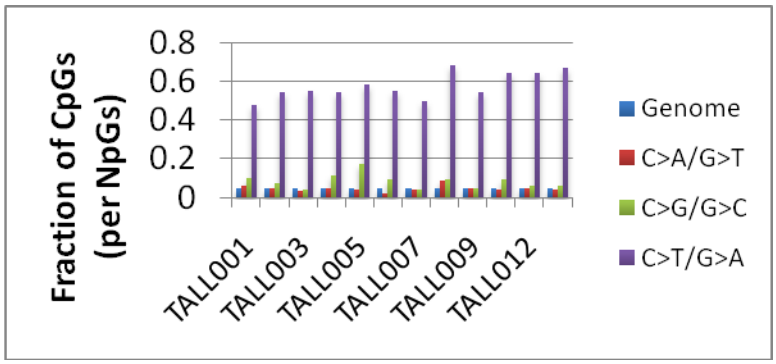
b


Number of Mutations
900
800
700
600
500
400
300
200
100
0
TALL001
TALL002
TALL003
TALL004
TALL005
TALL006
TALL007
TALL008
TALL009
TALL011
TALL012
TALL013
A>C/T>G
A>G/T>C
A>T/T>A
C>A/G>T
C>G/G>C
C>T/G>A

c

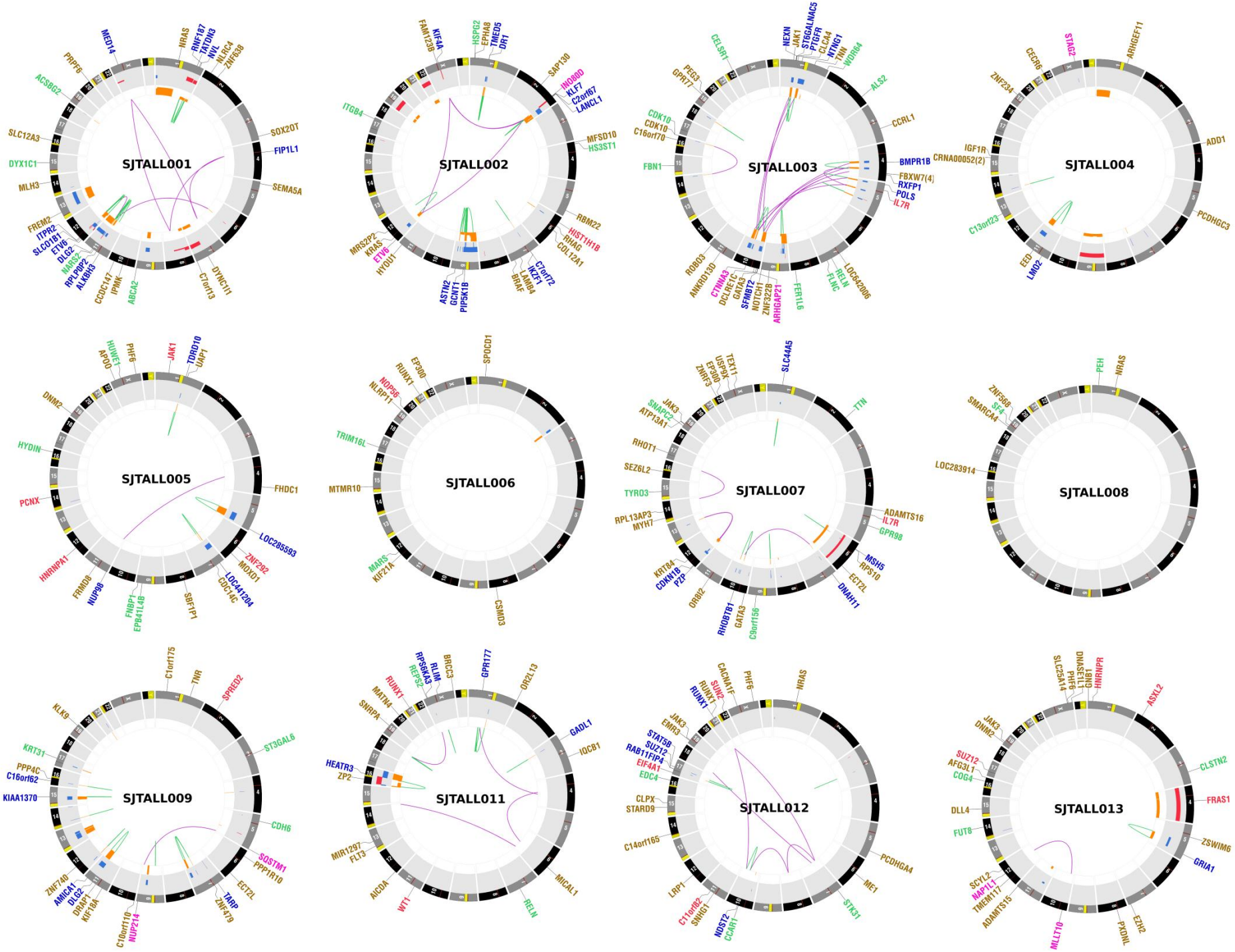

Percentage of Mutations
60
50
40
30
20
10
0
TALL001
TALL002
TALL003
TALL004
TALL005
TALL006
TALL007
TALL008
TALL009
TALL011
TALL012
TALL013
Sample
A>C/T>G
A>G/T>C
A>T/T>A
C>A/G>T
C>G/G>C
C>T/G>A

d

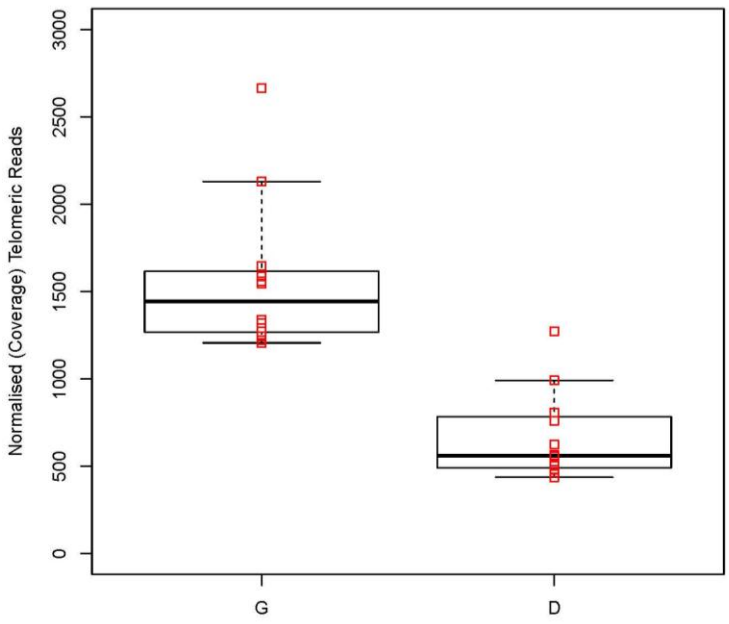

Fraction of CpGs (per NpGs)
0.8
0.6
0.4
0.2
0
TALL001
TALL003
TALL005
TALL007
TALL009
TALL012
Genome
C>A/G>T
C>G/G>C
C>T/G>A

Supplementary Figure 6. CIRCOS plots of genetic alterations in all 12 WGS cases

The CIRCOS[94] plots depicting structural genetic variants, including DNA copy number alterations, intra- and inter-chromosomal translocations, and sequence alterations. LOH, orange; amplification, red; deletion, blue; Sequence mutations in Refseq genes: silent single nucleotide variants (SNVs), green; non-silent SNVs, brown; indels, red; genes at structural variant breakpoints: genes involved in in-frame fusions, pink; others, blue.

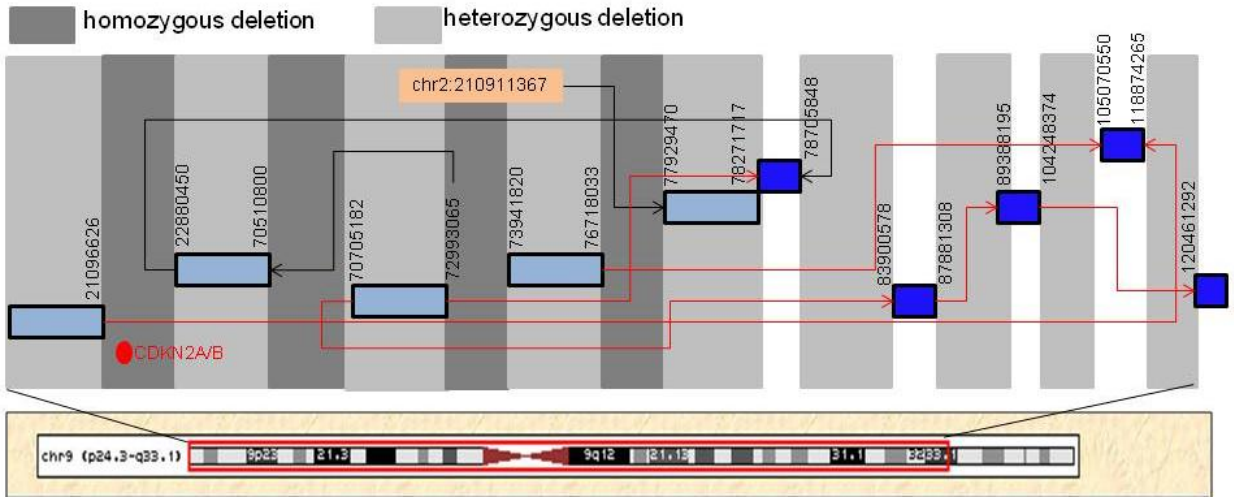

SJTALL001
SJTALL002
SJTALL003
SJTALL004
SJTALL005
SJTALL006
SJTALL007
SJTALL008
SJTALL009
SJTALL011
SJTALL012
SJTALL013

Supplementary Figure 7. Telomere shortening in ETP ALL



Telomere shortening has long been established in a number of cancers, and in most cancers telomere length is maintained above a critical threshold to ensure that the replicative capacity of the cancer cell is retained. The SJTALL dataset corroborates these findings, with a significant reduction (t-test *P*=0.00001) in the amount of normalised telomeric reads in the diagnostic samples, but not a total loss of telomeres with an average normalised read count in the diagnostic samples of 664.

Supplementary Figure 8. An example of multiple complex inter-chromosomal rearrangements in sample SJTALL002.

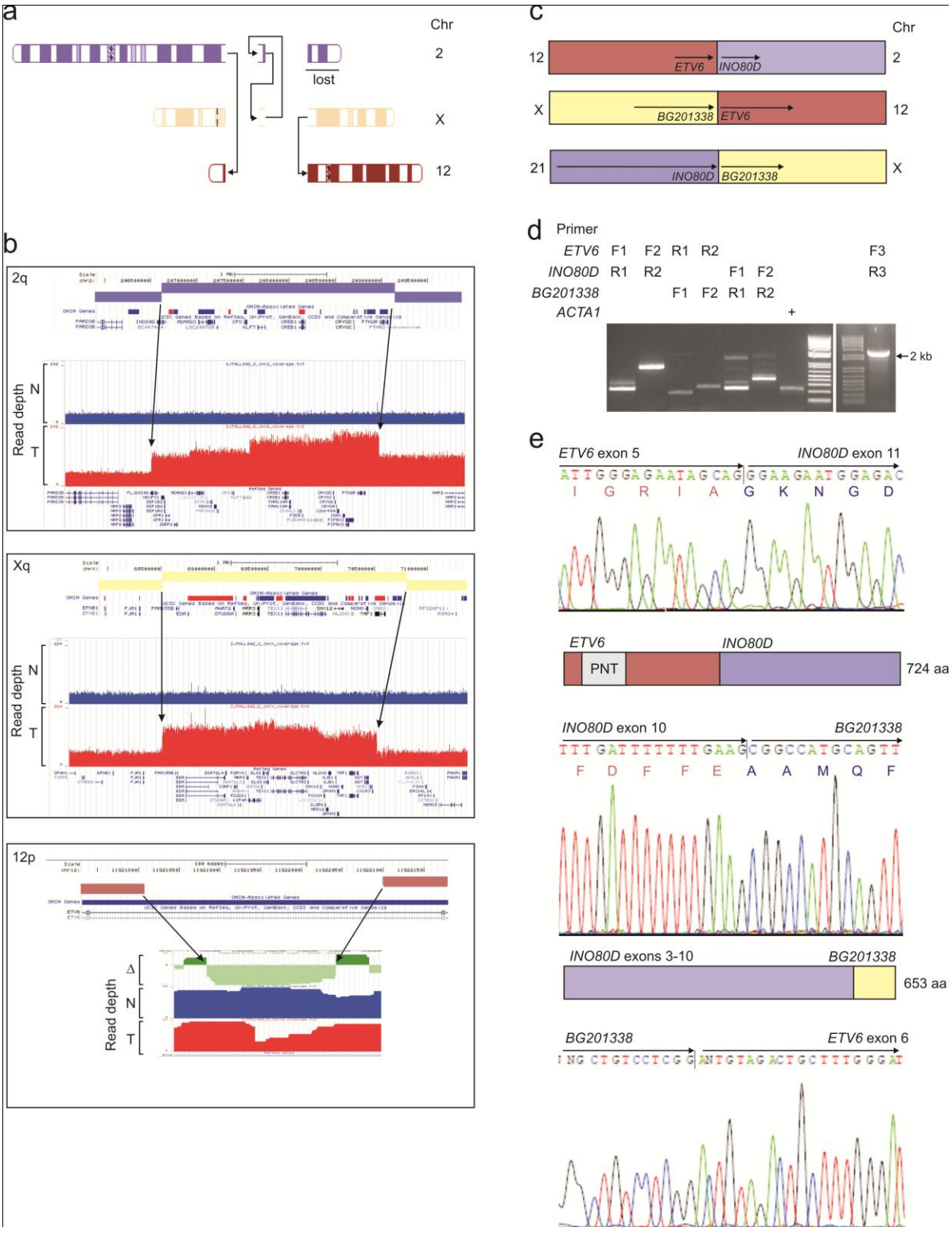


A total of 9 re-arrangements at chr9:1-120,461,292 resulted in 9 hemizygous deletions and 4 homozygous deletions (including homozygous deletion of *CDKN2A/CDKN2B*) in this region. Lines connecting the segments were based on the location and the orientation of structural variations computed by the CREST program. A red line indicates that the SV has been experimentally validated while a black line indicates the SV is unvalidated due to assay failure. The genomic location of each breakpoint is marked with base-pair resolution. The distance is not drawn in proportion to the size.

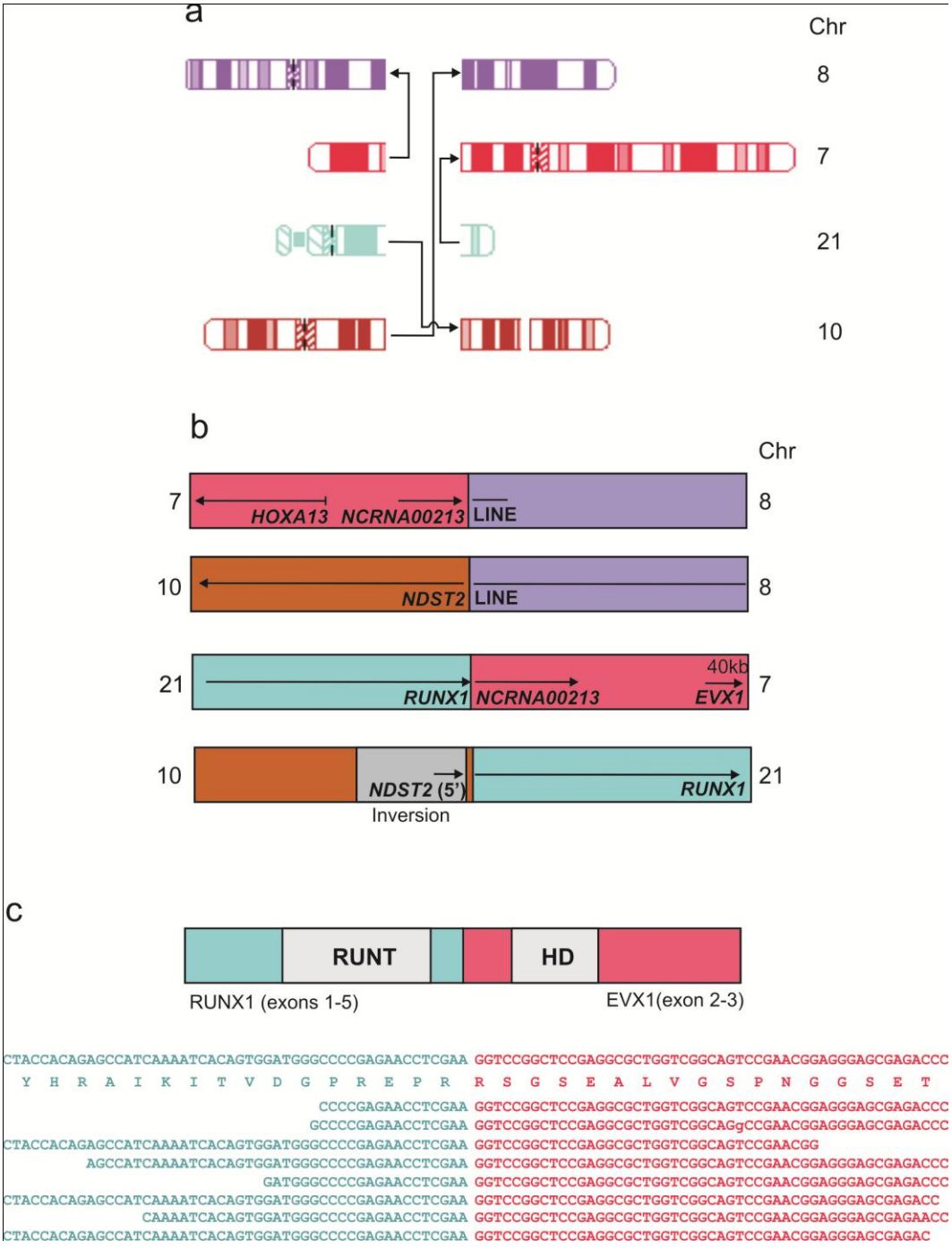
Supplementary Figure 9. Complex rearrangement of *ETV6* in SJTALL002.

a, the rearrangement involves chromosomes 2q33.3-34 (at *INO80D*), Xq13.1 (at *BG201338*) and 12p13.2 (at *ETV6*). **b**, amplifications adjacent to the breakpoints at 2q33 and Xq13.1, and a focal deletion within *ETV6*. **c-e**, The rearrangement results in expression of *ETV6-INO80D, INO80D-BG201338* and *BG201338-ETV6*, confirmed by RT-PCR (**d**) and Sanger sequencing (**e**). *ETV6-INO80D* is predicted to encode a 724 amino acid protein that retains the pointed domain of ETV6. *INO80D-BG201338* encodes a 653 amino acid protein of unknown function, and *BG201338-ETV6* leads to an out-of-frame fusion. The right panel of **d** indicates PCR and cloning of full-length *ETV6-INO80D.*

Supplementary Figure 9



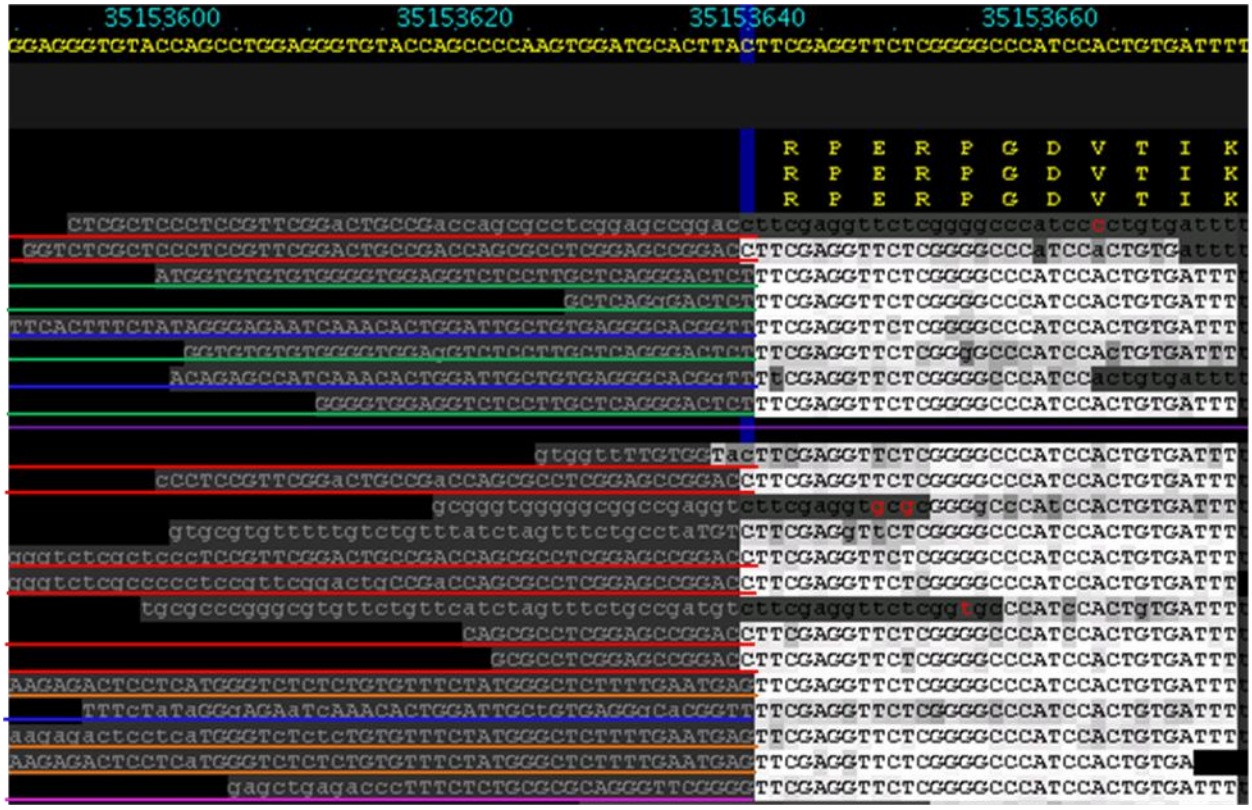
Supplementary Figure 10. Complex rearrangements in case SJTALL012.



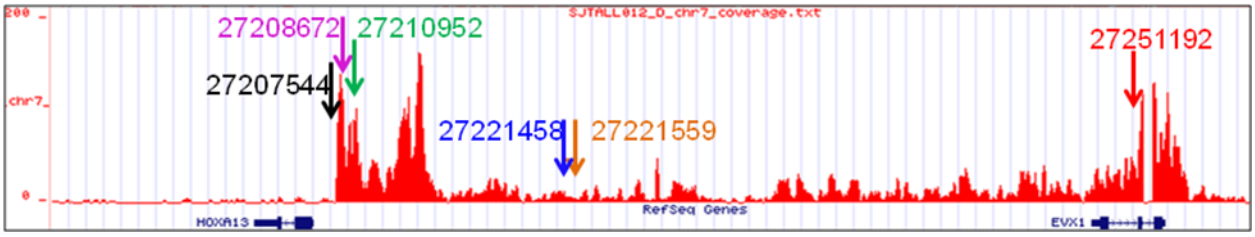
a, Depiction of a four way translocation involving chromosomes 7, 8, 10 and 21; and **b**, putative partner gene rearrangements arising from the translocation. The translocation of chromosome 21 to chromosome 7 results in juxtaposition of the 5' region of *RUNX1* to the noncoding RNA gene, *NCRN00213*, approximately 40kb upstream of the homeobox gene *EVX1.* **c**, Transcriptome sequencing demonstrates that this rearrangement results in expression of a chimeric transcript that joins exons 1-5 of *RUNX1* to exons 2-3 of *EVX1.*

Supplementary Figure 11. RNA-seq identifies RUNX1 rearrangement in SJTALL012.

a

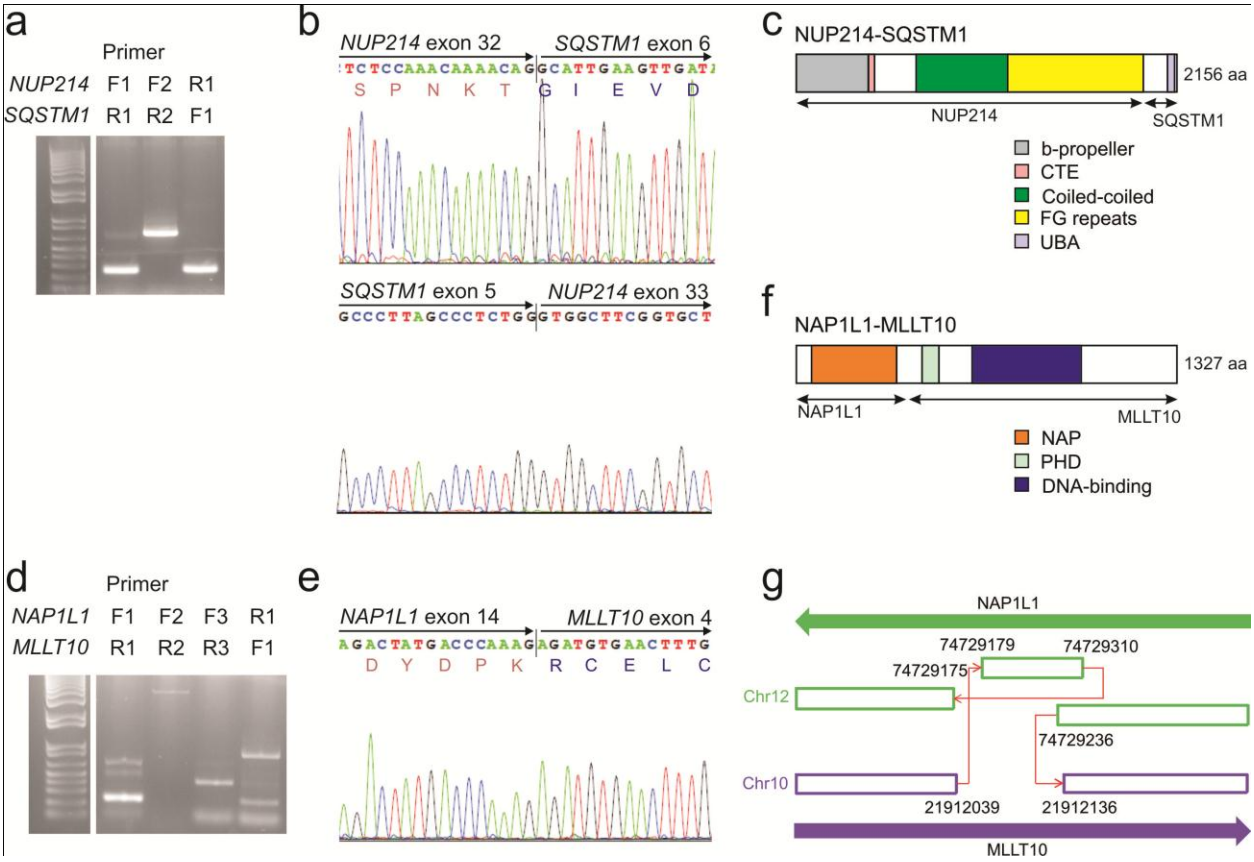


b



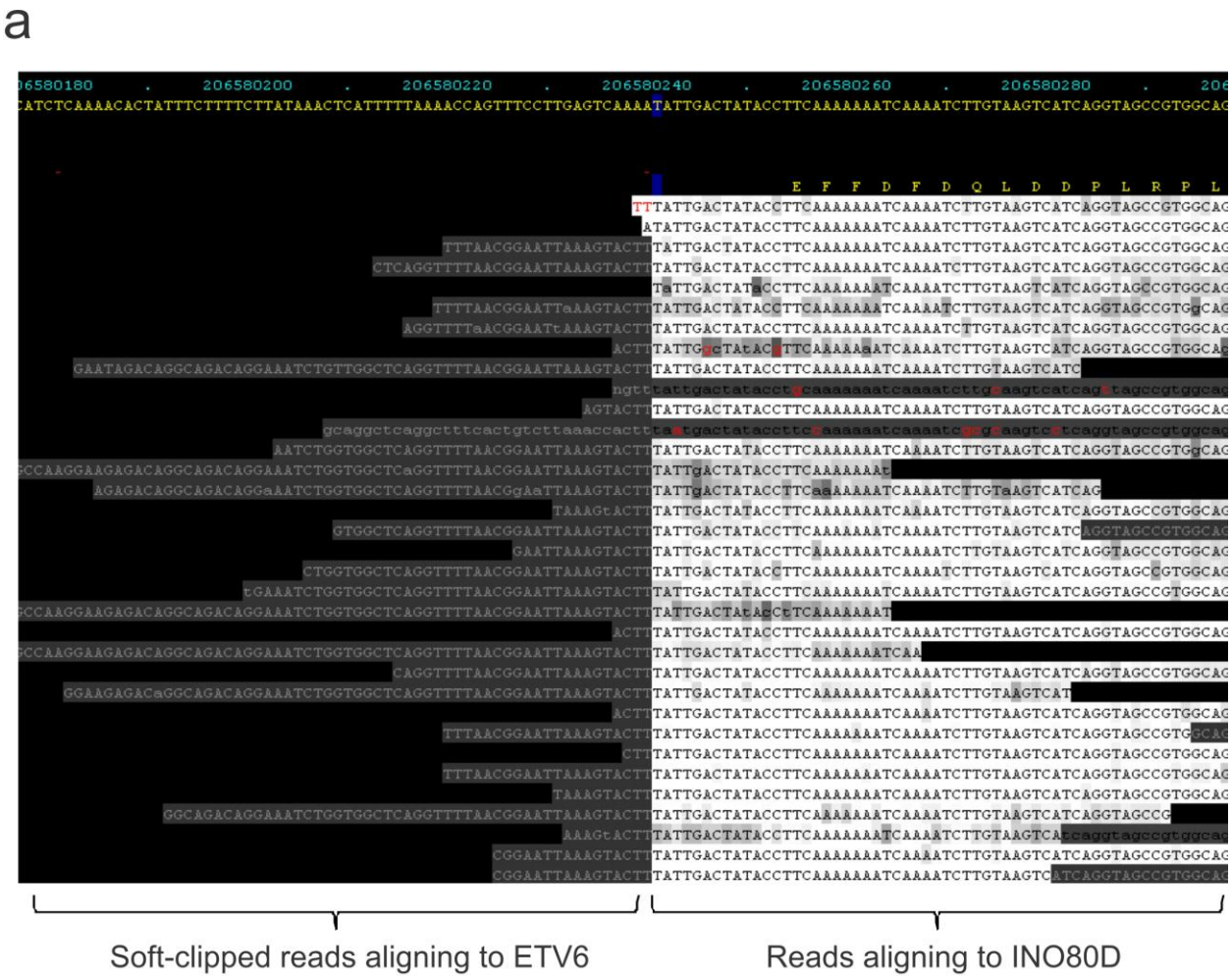
RNA-seq identifies multiple fusion partners of *RUNX1* exon 5 caused by aberrant splicing events and read-through from the *RUNX1* translocation breakpoint. **a,** Soft-clipped reads that represent a distinct partner are underlined with the same colour for each partner. The red underline represents the *RUNX1-EVX1* in-frame fusion protein. **b,** Genomic location of the *RUNX1* fusion partners. The breakpoint of the genomic DNA is labelled in black while the other colours match the fusion partners labelled in **a**. The vertical height of the red bar indicates the number of high-quality reads covering the site. The maximum height is 200 (i.e. 200 reads)

Supplementary Figure 12. Reciprocal inter-chromosomal translocations in ETP ALL.



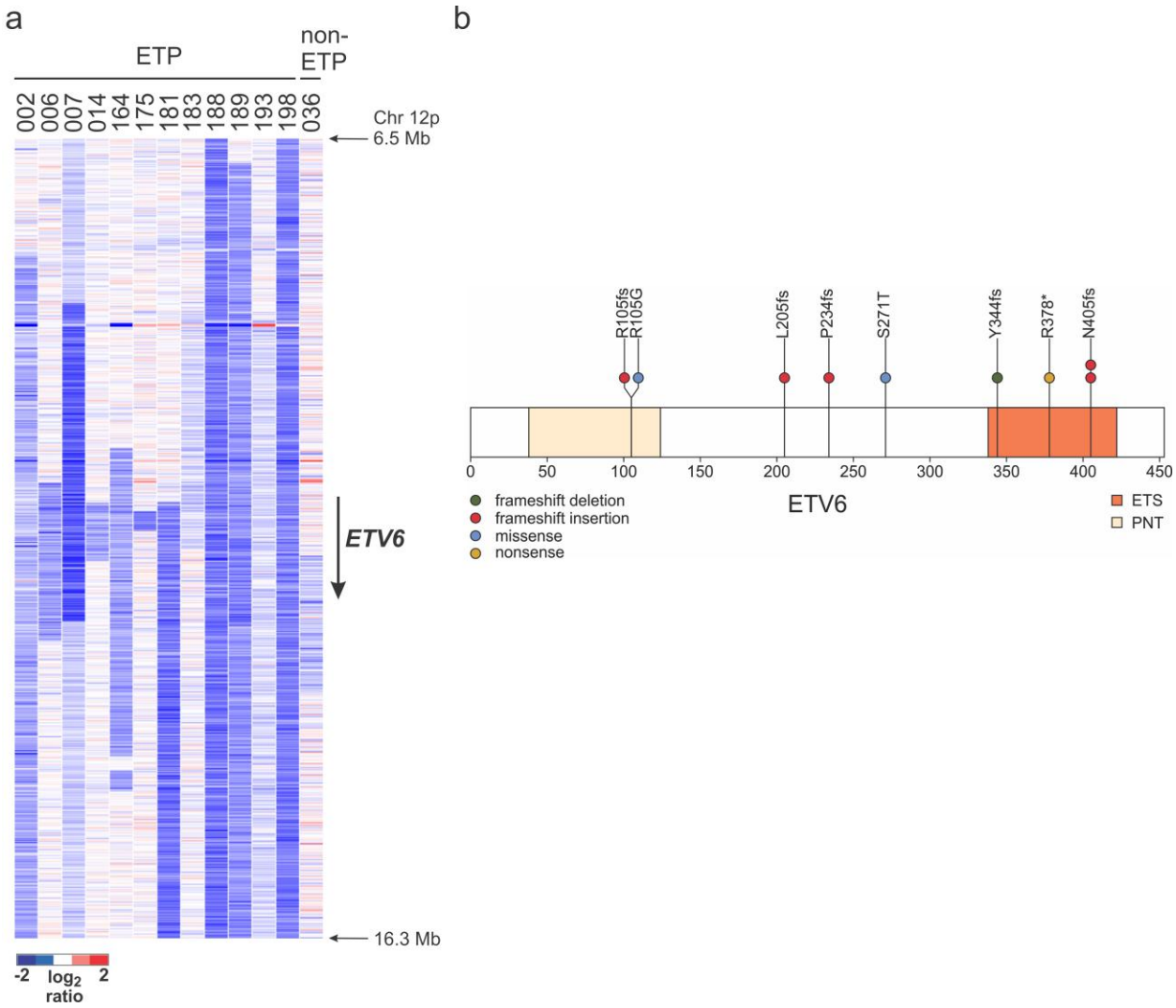
a-c, The *NUP214-SQSTM1* rearrangement in SJTALL009. **a**, RT-PCR; **b**, Sanger sequencing of RT-PCR products; and **c**, predicted domain structure of the chimaeric fusion protein. **d-g**. *NAP1L1-MLLT10* rearrangement in SJTALL013. **d**, RT-PCR; **e**, Sanger sequencing of RT-PCR products, and **f**, predicted domain structure of the chimaeric fusion protein. **g**, Genomic structure of the inter-chromosomal translocation (CTX) in SJTALL013 that involves chromosomes 10 and 12 with a 131bp inversion which occurs on the same haplotype as the CTX. The transcription orientation for *NAP1L1* and *MLLT10* are marked accordingly.

Supplementary Figure 13. Detection of *ETV6-INO80D* in case SJTALL208 by whole exome sequencing.



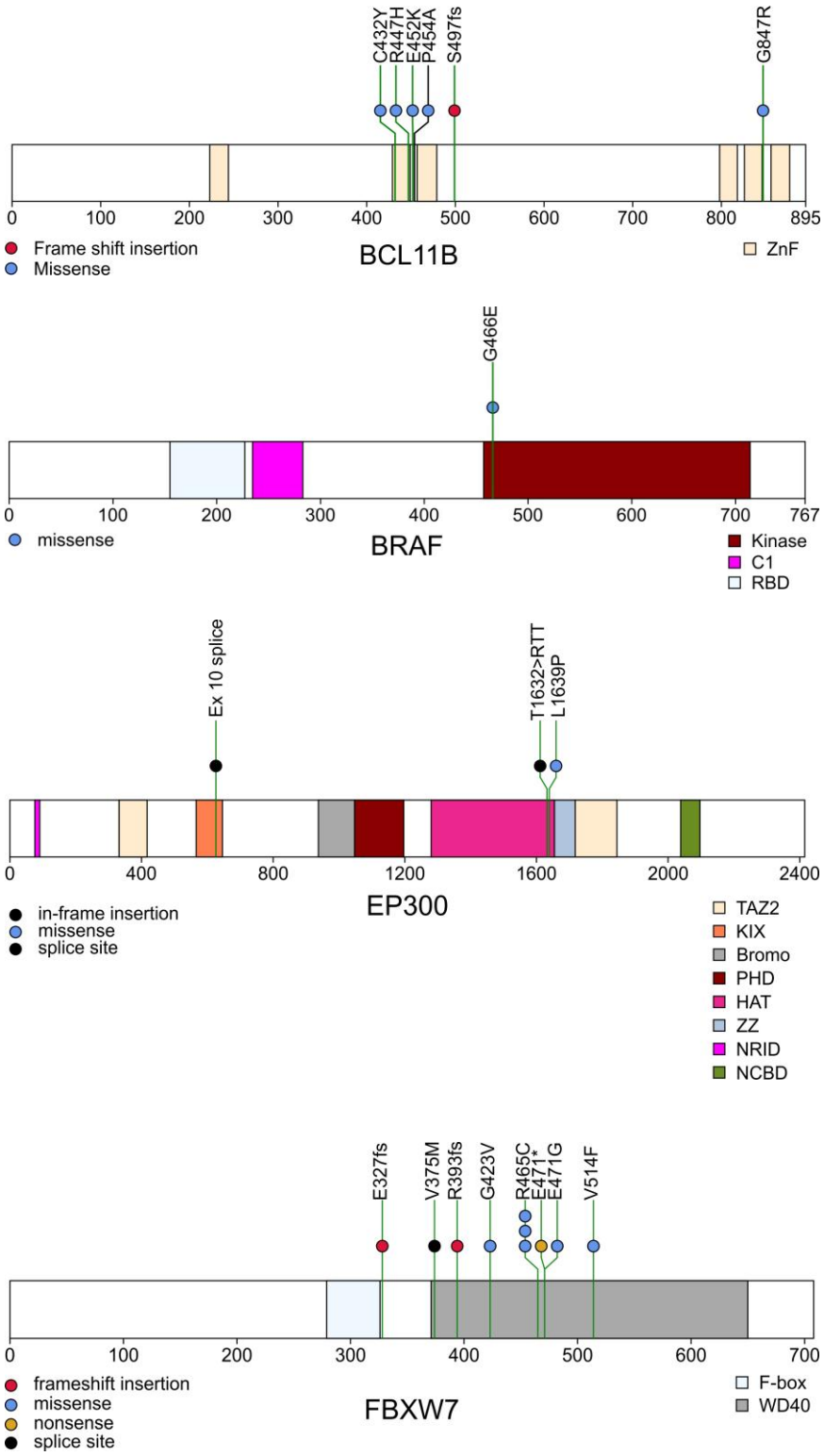
a, Soft-clipped reads aligning to *ETV6* and *INO80D* indicating the presence of the ETV6-INO80D rearrangement, confirmed by genomic PCR (**b**) and direct Sanger sequencing of the PCR product (**c**).

Supplementary Figure 14. *ETV6* deletions and sequence mutations in ETP ALL.

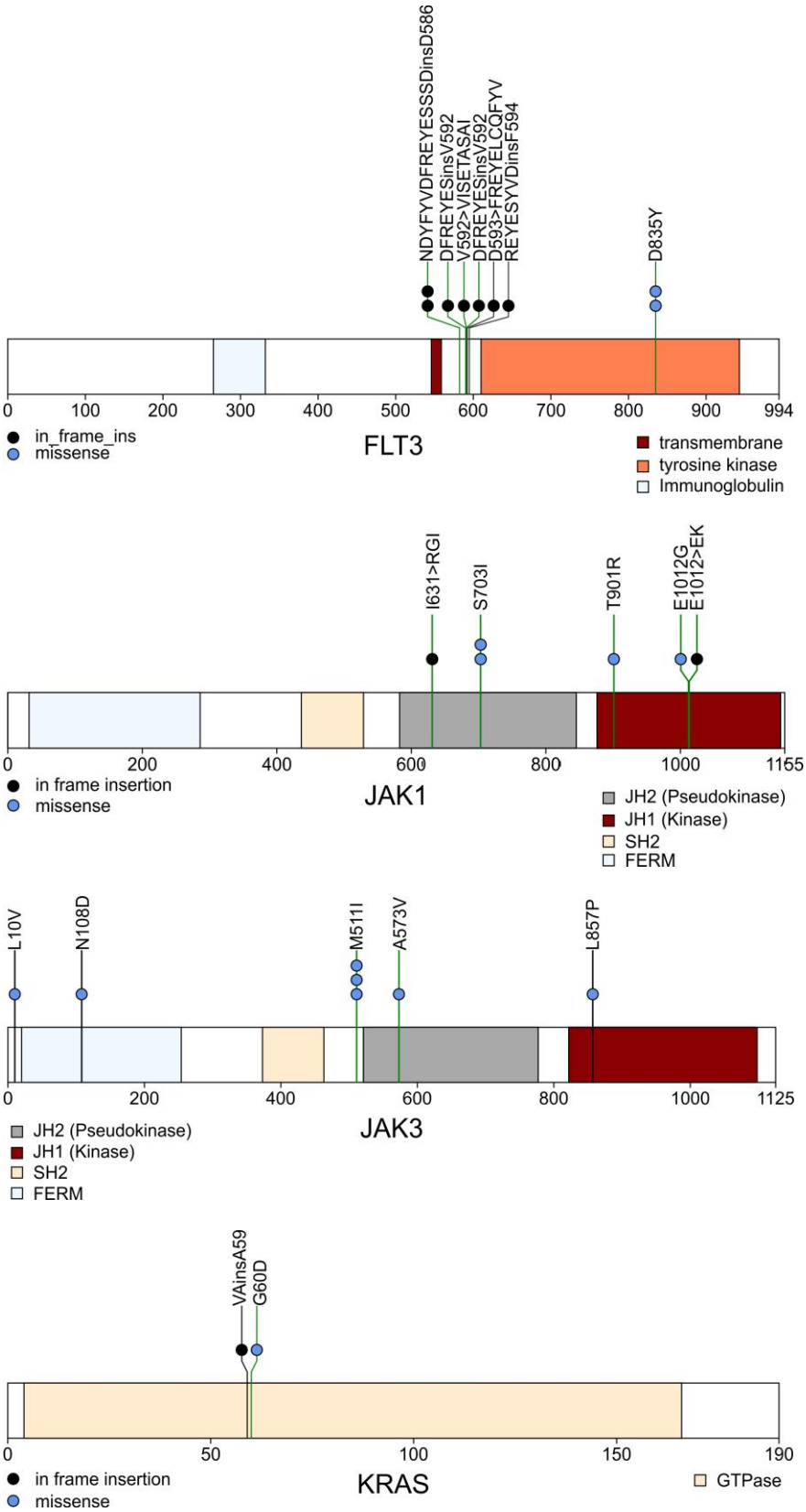


a, Thirteen cases harboured focal or broad deletion of *ETV6*. Representative SNP 6.0 microarray data is shown. **b**, cases with ETV6 sequence mutations.

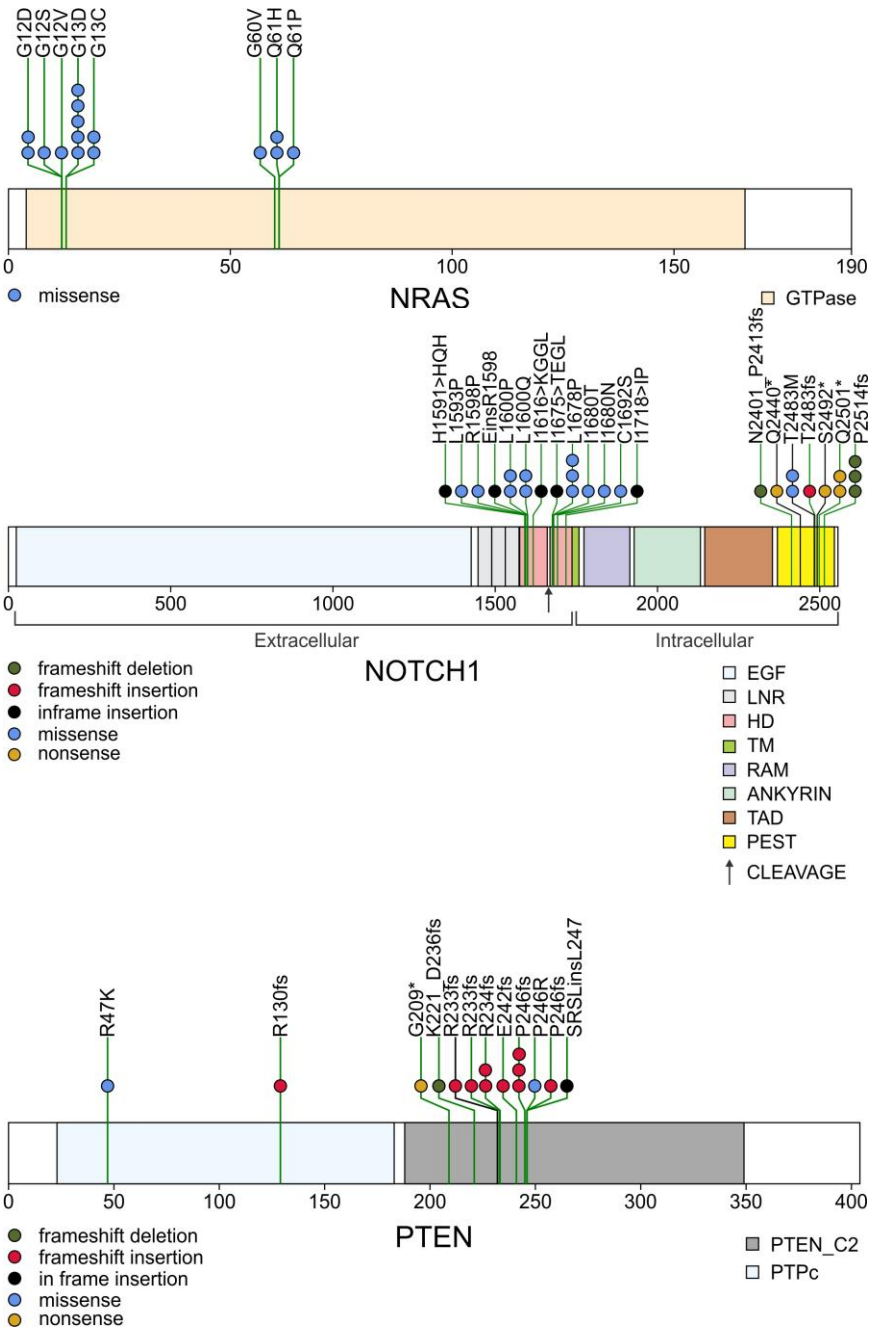
Supplementary Figure 15. Protein domain and mutation plots for targets of recurring and novel sequence mutation in T-ALL.



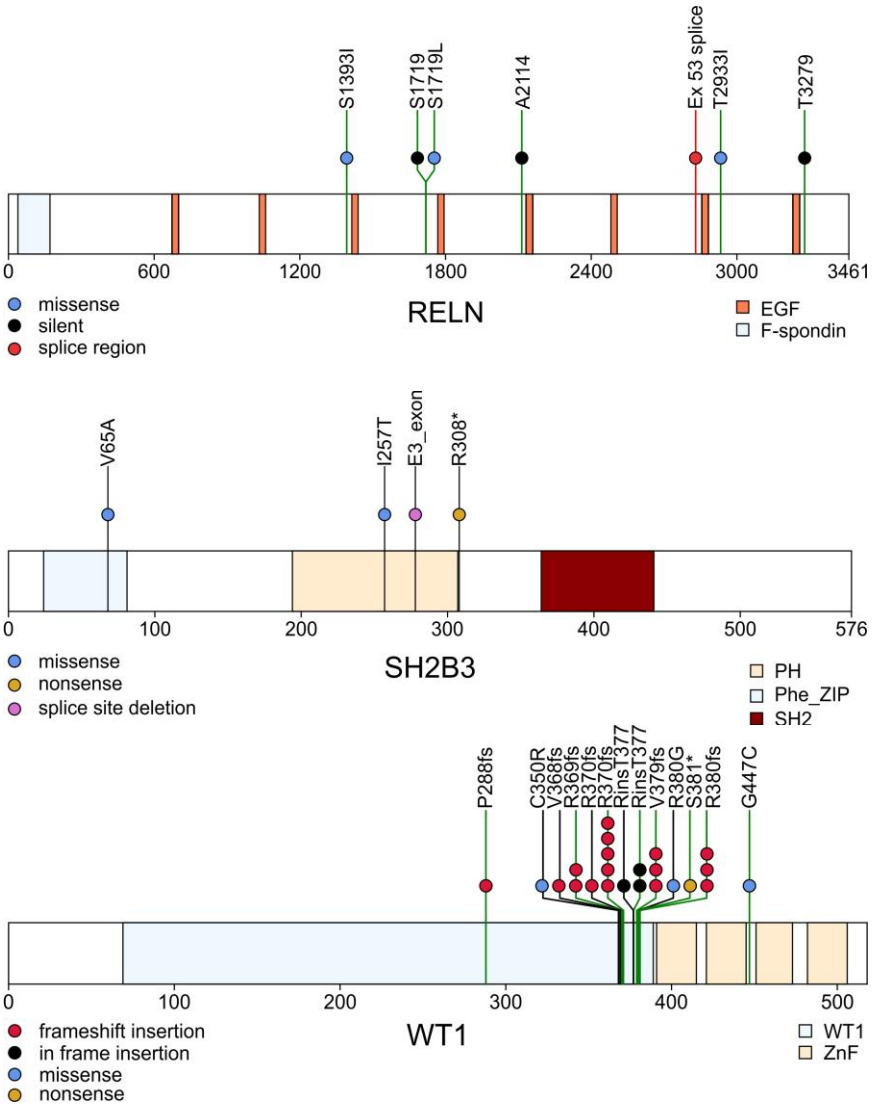
Supplementary Figure 15 (continued)



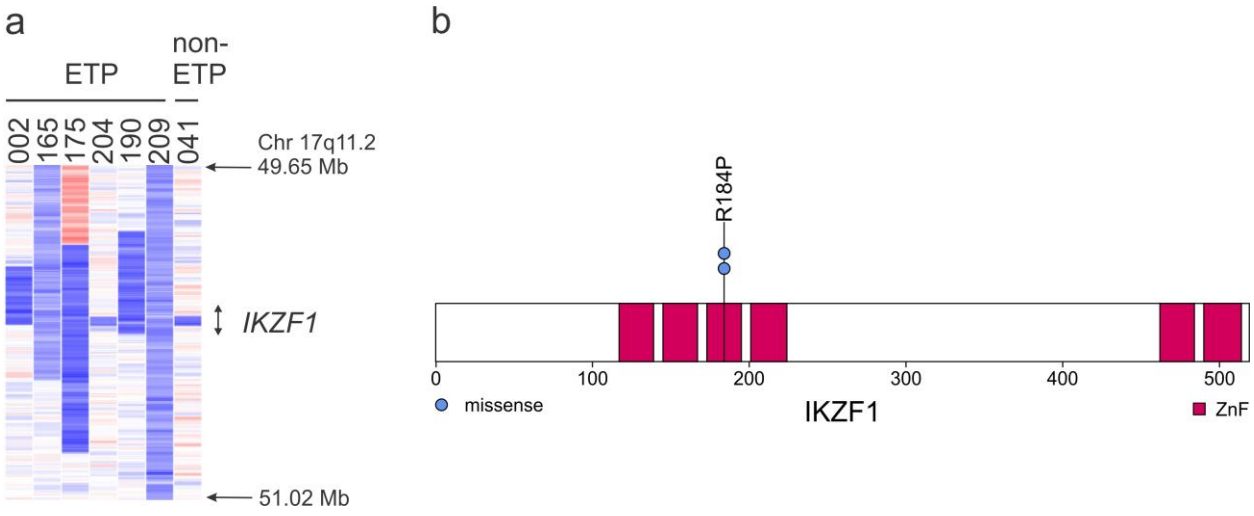
Supplementary Figure 15 (continued)



Supplementary Figure 15 (continued)

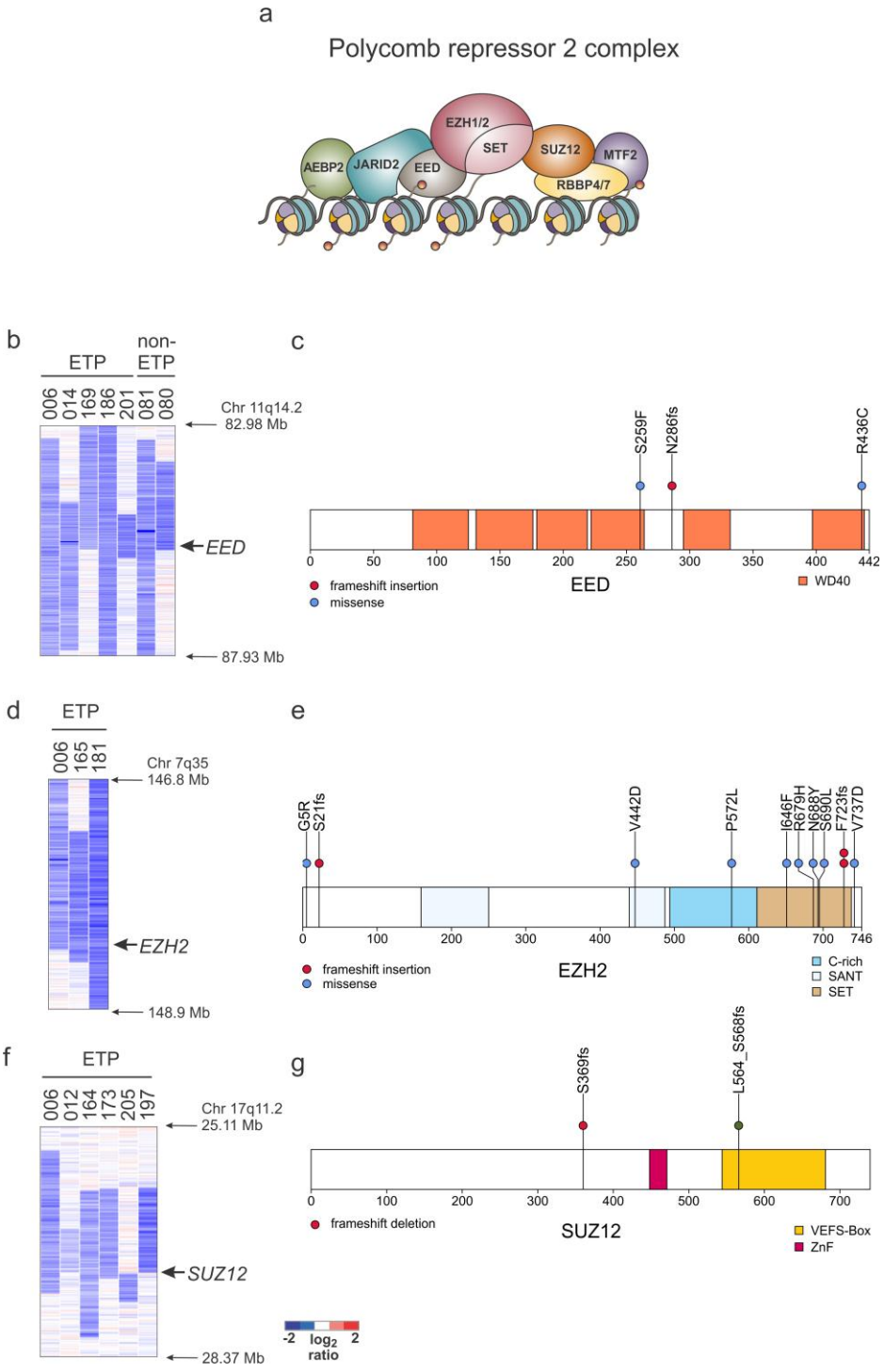


Supplementary Figure 16. Deletions and mutations of *IKZF1* (IKAROS) in ETP ALL.



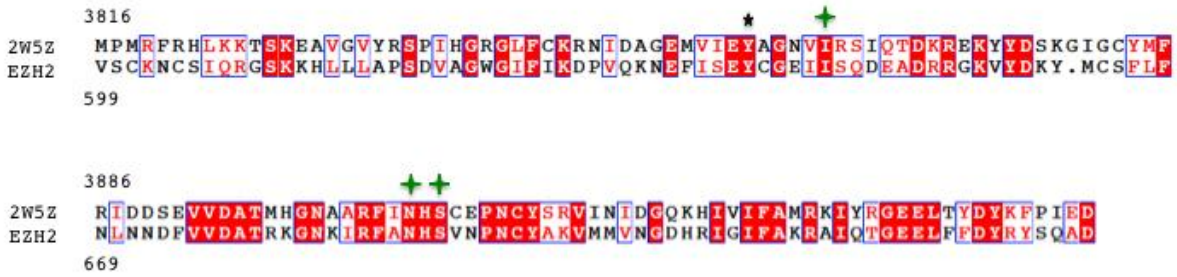
a, Affymetrix SNP 6.0 data showing focal and broad deletions of *IKZF1* in T-ALL. Cases SJTALL204 and SJTALL041 have a deletion of exons 4-7 (coding exons 3-6) that results in expression of a dominant negative isoform of IKZF1, IK6, that is commonly observed in *BCR-ABL1* positive[5] and high risk *BCR-ABL1* negative B-progenitor ALL.[34] **b**, sequence mutations in IKZF1. R184 is located at a region of the third zinc finger of IKZF1 that is critical for DNA binding.[95]

Supplementary Figure 17. Polycomb repressor 2 complex mutations in ETP T-ALL



a, Schematic of components of the polycomb repressor complex 2 (PRC2) that mediates histone 3 lysine 27 (H3K27) trimethylation. Schematic adapted from ref. [96]. **b-g**, Affymetrix SNP 6.0 DNA log2 ratio plots showing examples of PRC2 gene deletions, and sequence mutations identified by WGS and Sanger sequencing.

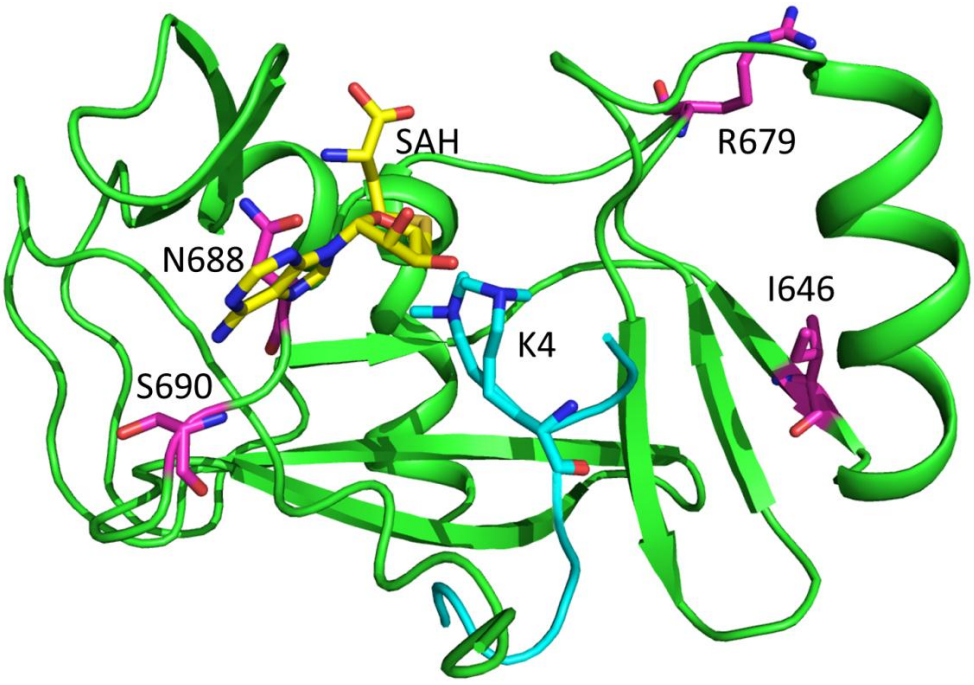
Supplementary Figure 18. Sequence alignment of the EZH2 SET domain with the sequence of the structurally characterised MLL1 SET domain (PDB: 2W5Z).



Identical residues are enclosed in boxes with a red background and similar residues are in red type and enclosed in boxes with a white background. Residues identified as missense mutations (I646F, R679H, N688Y, and S690L) are indicated by a green plus symbol above the position in the sequence and the previously identified mutation site, Y641, is indicated by a black asterisk.

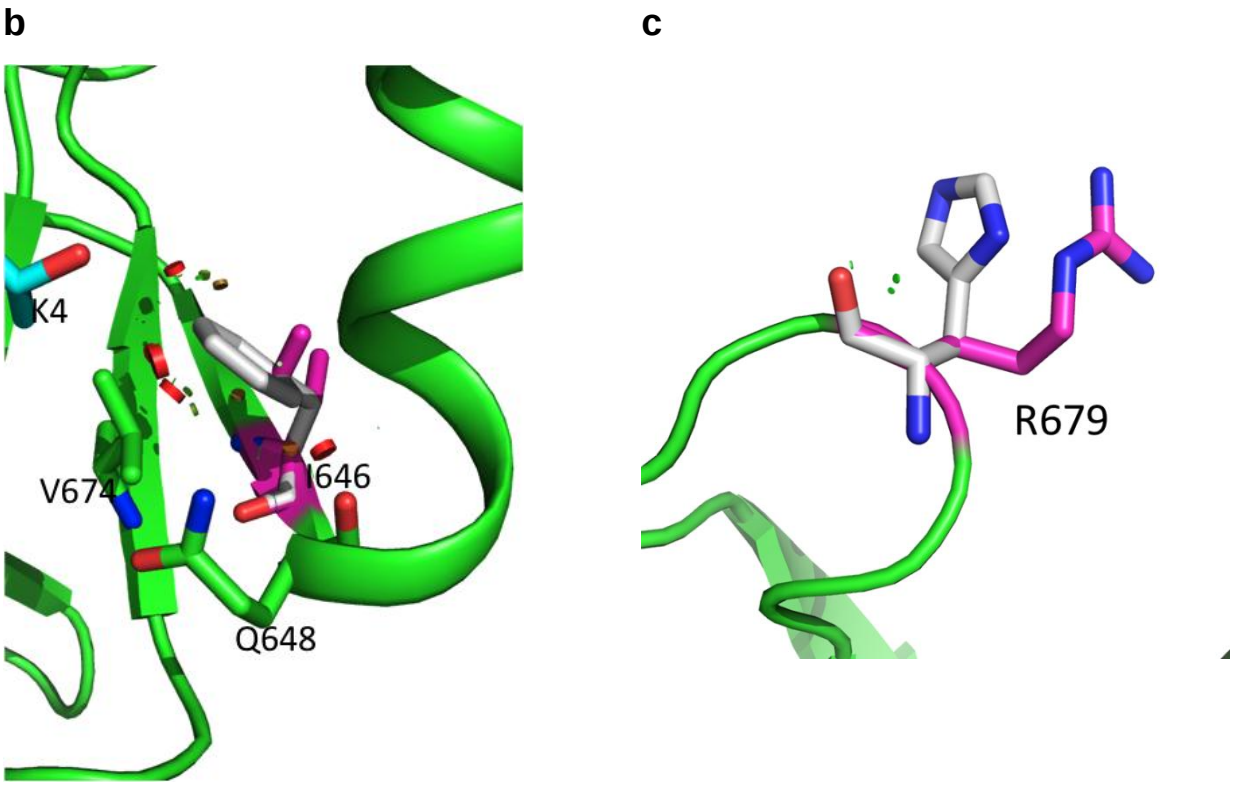
Supplementary Figure 19. Structural modelling of EZH2 mutations

a



a, Ribbon representation of the EZH2 SET domain homology model with amino acids that are mutated in T-ALL patients illustrated with purple sticks. Stick representations of the S-Adenosylhomocysteine cofactor (SAH: yellow) and a methylated lysine-containing substrate peptide [Lys 4 (K4: teal)] from the reference MLL1 SET domain structure (PDB: 2W5Z) are also shown. In the reference structure, Lys 4 exhibited two conformations and is modelled in this way here.

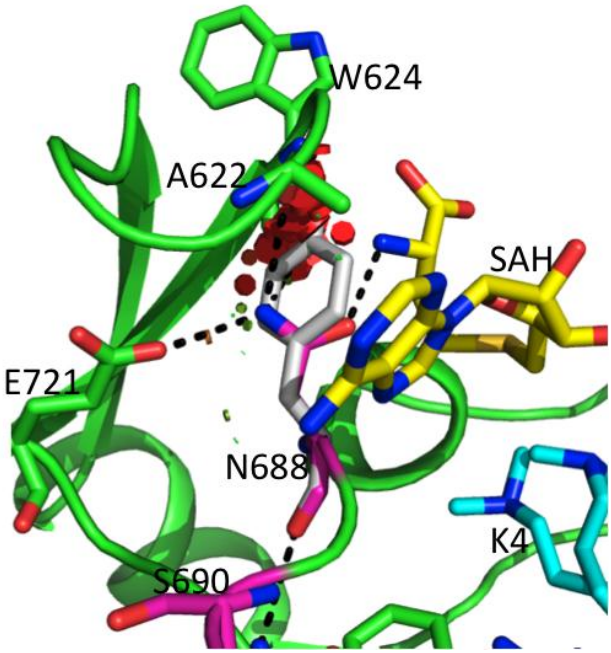
Supplementary Figure 19 (continued)



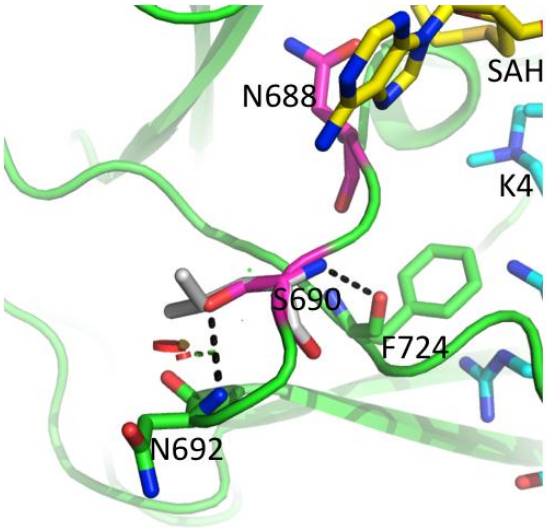
b, Ribbon representation of the region of the EZH2 SET domain model near Ile 646 (I646; illustrated with purple sticks) and stick representation of selected residues within 4 Å (illustrated with green sticks). Substitution of Ile 646 (I646) with Phe (illustrated with white sticks) creates small steric clashes with Val 674 (V674) and Gln 648 (Q648; clashes indicated by red disks). **c,** Ribbon representation of the region of the EZH2 SET domain model near residue Arg 679 (R679; illustrated with purple sticks). A substitution of Arg679 with His (illustrated with white sticks) does not create steric clashes with adjacent residues.

Supplementary Figure 19 (continued)

d



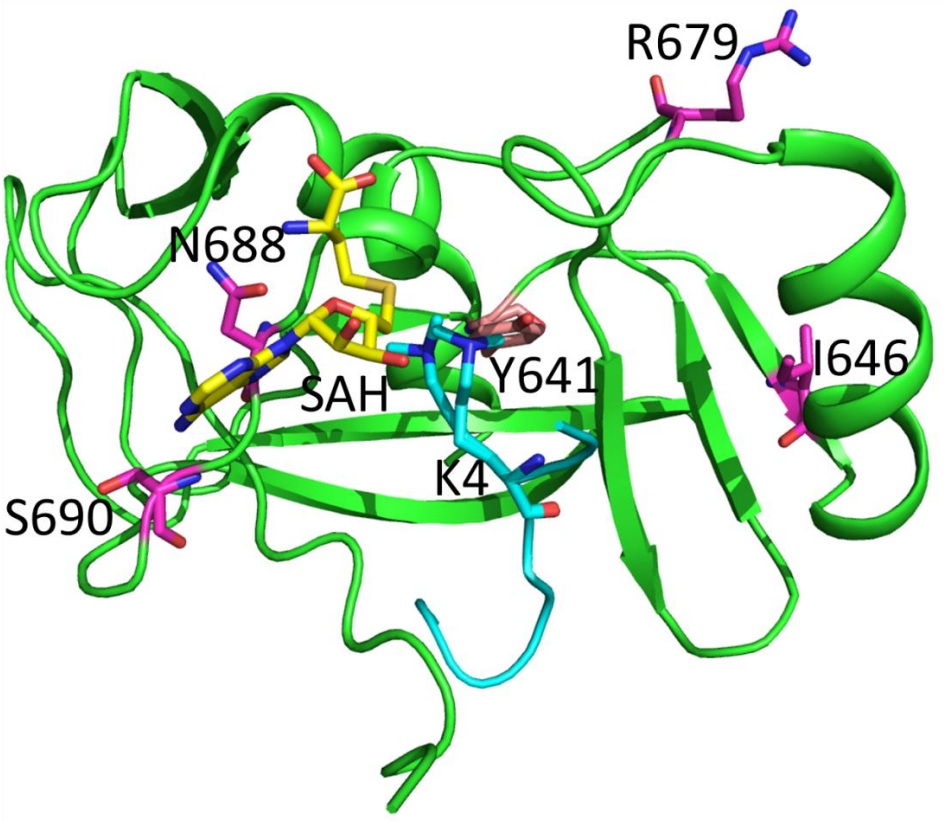
e



d, Ribbon representation of the region of the EZH2 SET domain model near Asn 688 (N688; illustrated with purple sticks) and stick representation of selected residues capable of making interactions with Asn 688 within 4 Å [Ser 690 (S690) is illustrated with purple sticks and all others are illustrated with green sticks]. Potential polar interactions between side chain atoms are indicated with dotted lines. Substitution of Asn 688 with Tyr (illustrated with white sticks) creates steric clashes with residues within an adjacent loop (clashes indicated by red disks). Stick representations of the S-Adenosylhomocysteine cofactor (SAH: yellow) and a portion of a methylated lysine-containing substrate peptide (K4: teal) from the reference MLL1 SET domain structure (PDB: 2W5Z) are also shown. **e,** Ribbon representation of the region of the EZH2 SET domain model near Ser 690 (S690; illustrated with purple sticks) and selected residues capable of making interactions to Ser 690 within 4 Å [Asn 688 (N688) is also illustrated with purple sticks and all others are illustrated with green sticks). Potential polar interactions between side chain atoms are indicated with dotted lines. Substitution of Ser 690 with Leu (illustrated with white sticks) creates a small steric clash (clashes indicated by red disks) and results in the loss of a polar interaction. Stick representations of a portion of the S-Adenosylhomocysteine cofactor (SAH: yellow) and a portion of a methylated lysine-containing substrate peptide (K4: teal) from the reference MLL1 SET domain structure (PDB: 2W5Z) are also shown.

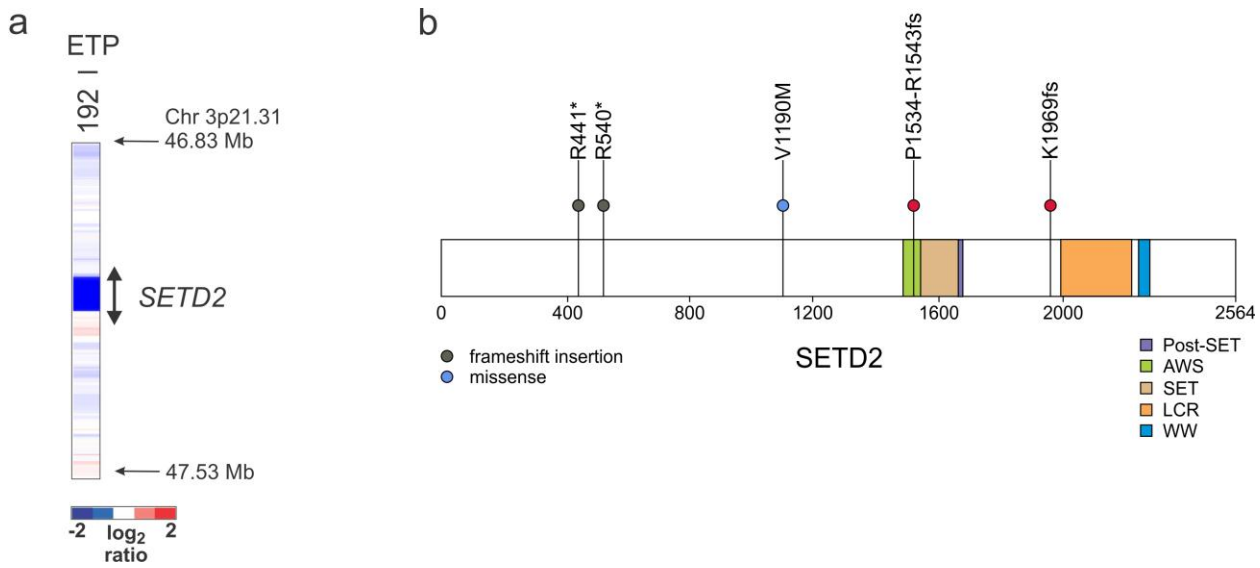
Supplementary Figure 19 (continued)

f

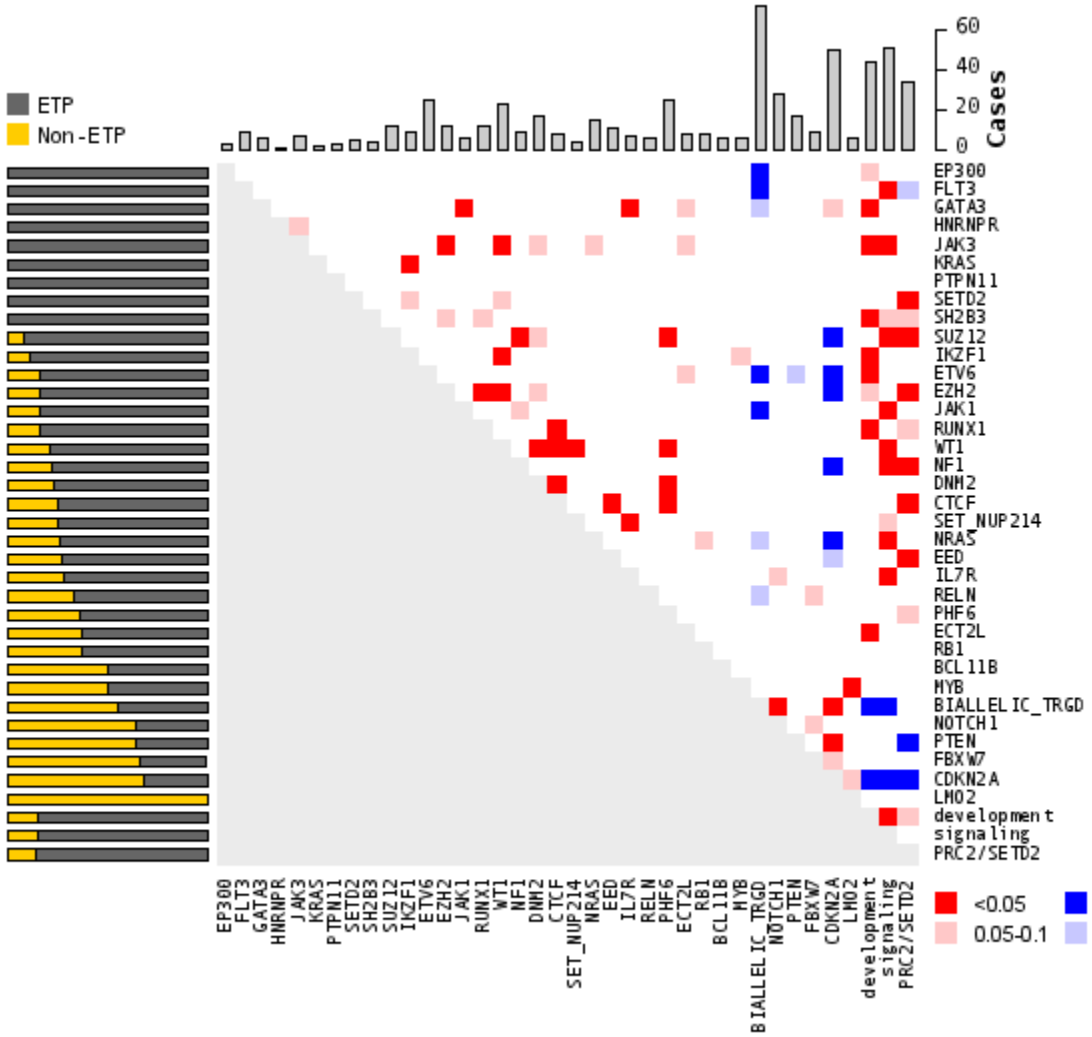


f, Ribbon structural view of the EZH2 SET domain model illustrating Tyr 641 (Y641) that was shown to be mutated in follicular and diffuse large B-cell lymphomas by Morin et al., (Tyr 641 is illustrated with peach sticks). Also illustrated are residues shown to be mutated in T-ALL patients in this study, including Ile 646 (I646), Arg 679 (R679), Asn 688 (N688) and Ser 690 (S690; each illustrated with purple sticks). Eight residues of the histone peptide substrate (teal) and the S-Adenosylhomocysteine cofactor (yellow) are also illustrated.

Supplementary Figure 20. *SETD2* alterations in ETP ALL

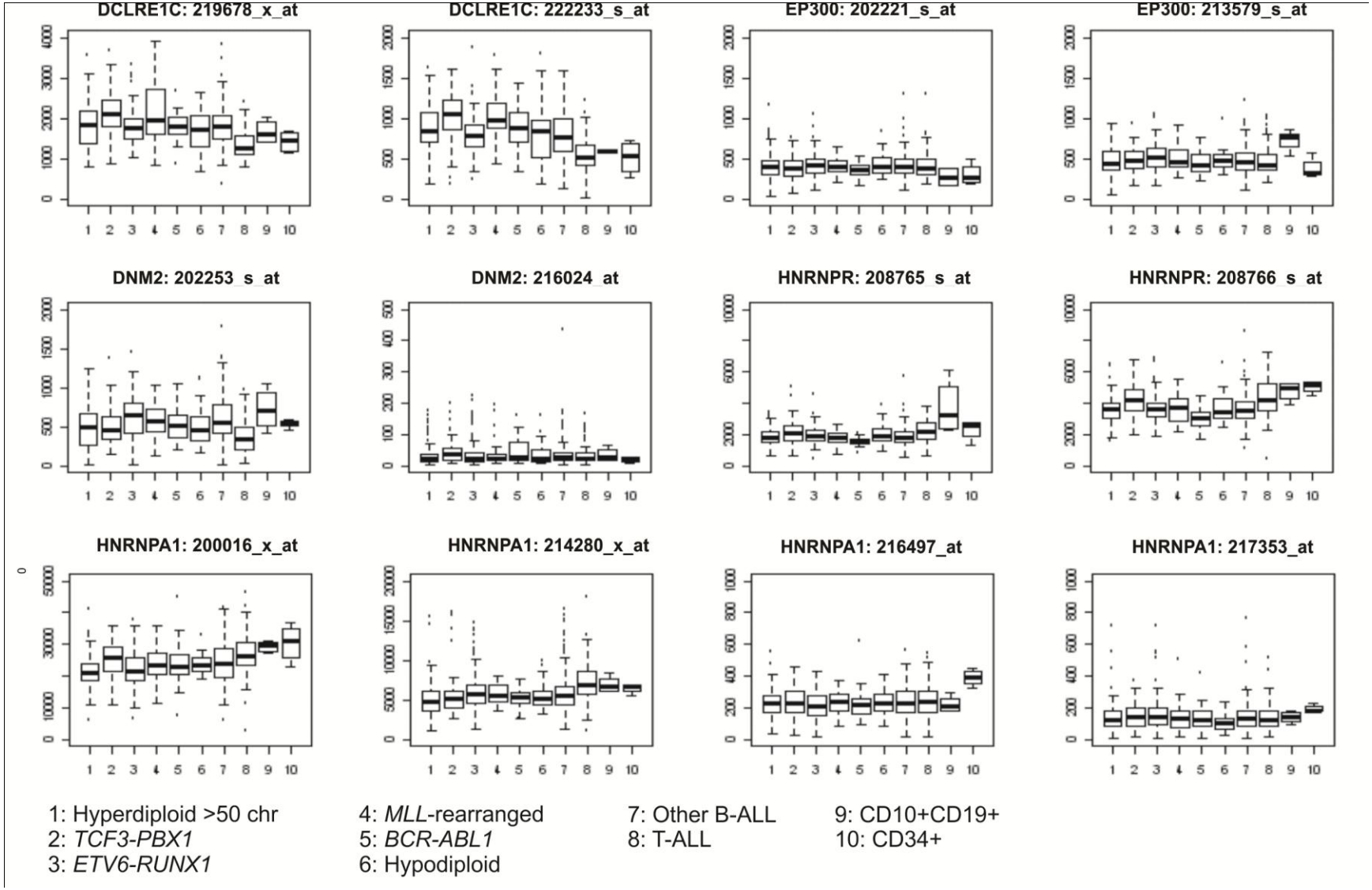


Supplementary Figure 21. Frequency and association lesion matrix.

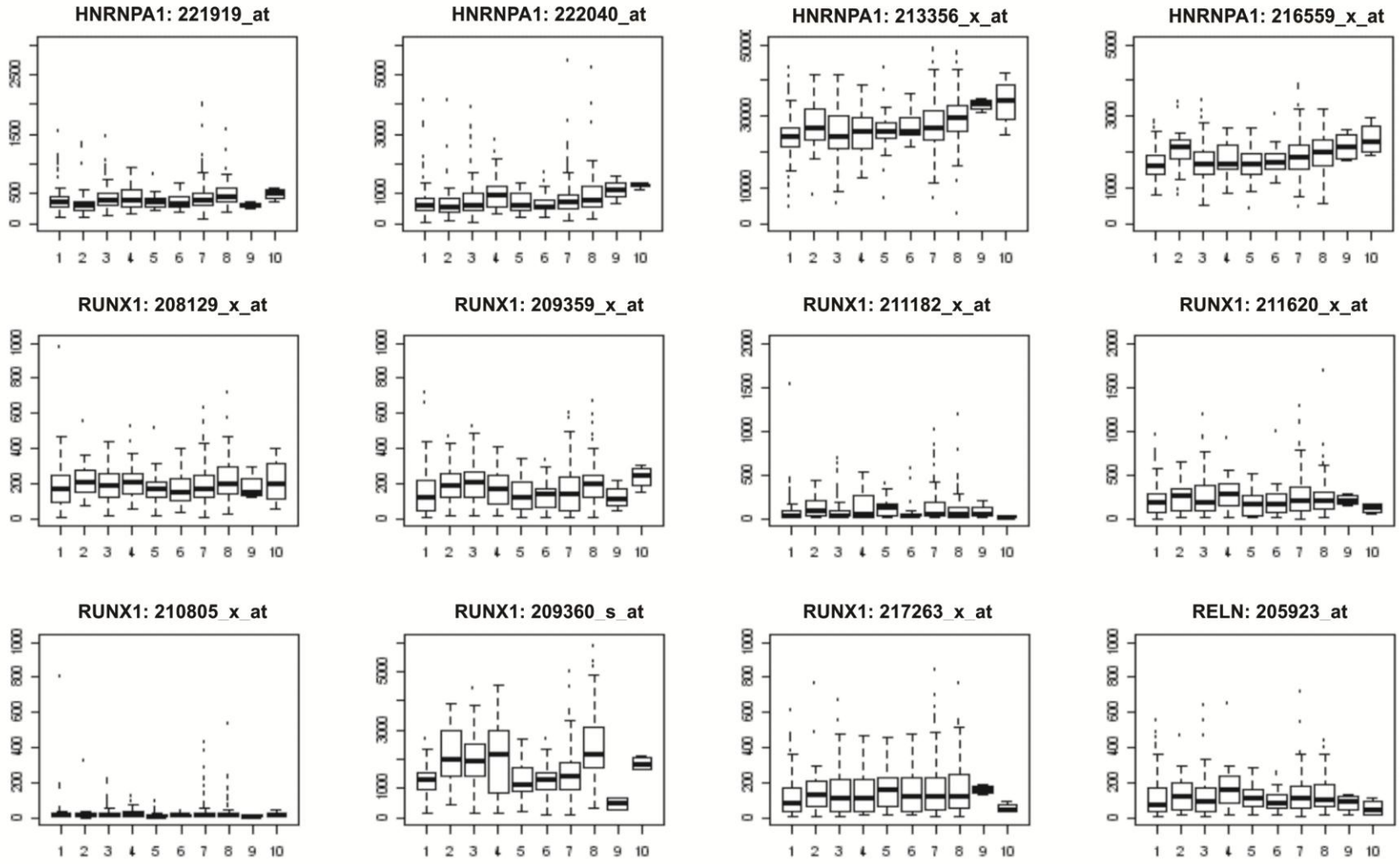


The figure depicts the frequency of recurring genetic alterations, and pathway lesions, and shows associations with ETP status, and associations between lesions. The top panel shows the number of cases harbouring an alteration in the genes as listed on the x axis. The left panel depicts the proportion of cases with an individual alteration that are ETP or non-ETP (i.e. all EP300, FLT3, GATA3, HNRPR and JAK3 lesions are in ETP cases, and all LMO2 alterations are in non-ETP cases). The large central panel shows associations between lesions (red: positive association; blue, negative association). Schematic adapted from Morin et al.[26].

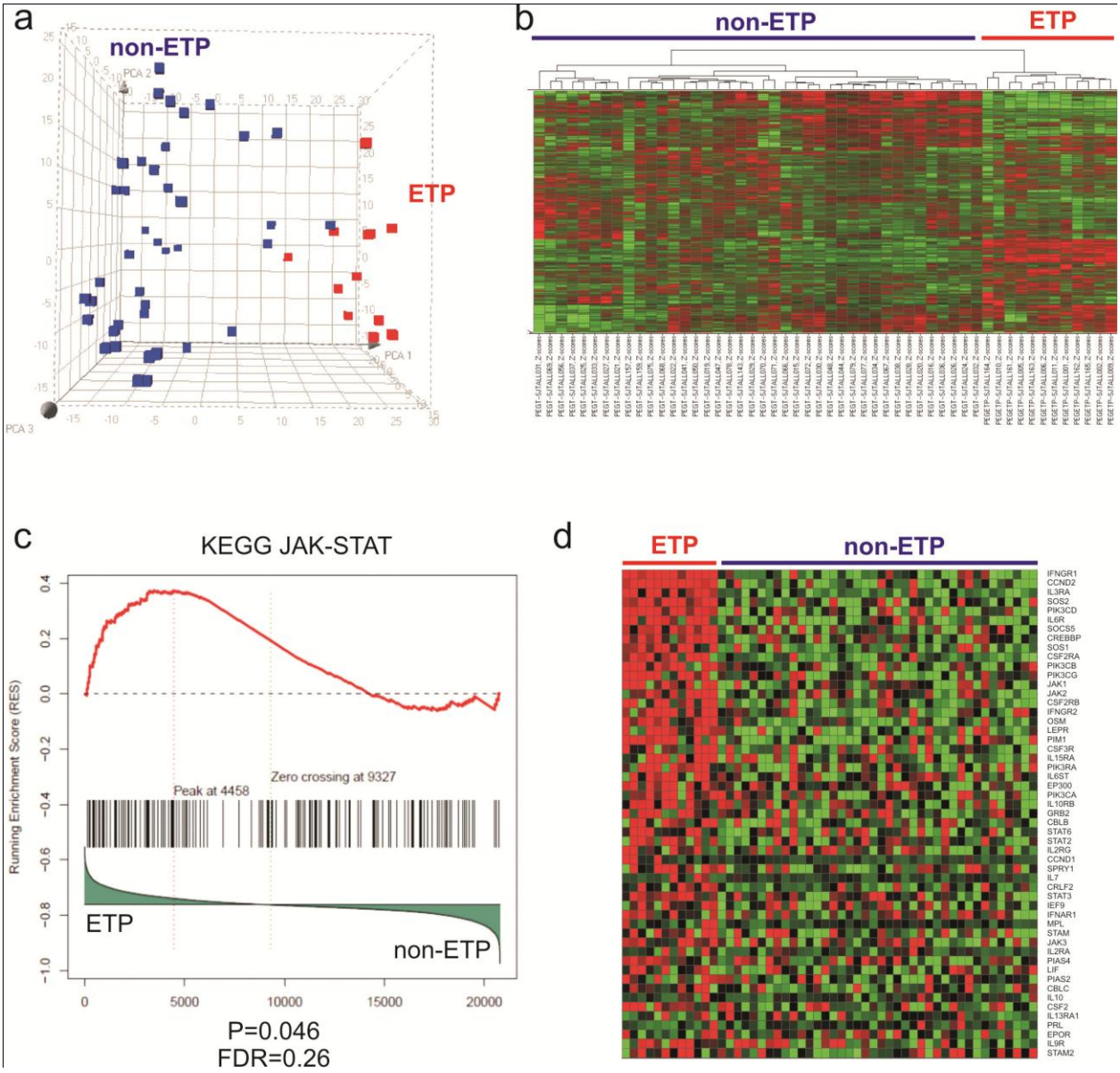
Supplementary Figure 22. Affymetrix U133A data showing box plots of expression levels of genes targeted by recurring sequence alterations in T-ALL.



Supplementary Figure 22 (continued)



Supplementary Figure 23. Gene expression profiling analysis of ETP ALL.



a, Principal components analysis using 1000 representative genes selected by k-means distinguishes ETP from non-ETP T-ALL. **b,** Unsupervised hierarchical clustering (Ward's method, 35074 probe sets) distinguishes ETP from non-ETP T-ALL cases. The signature of ETP ALL defined by *limma* analysis of the cases profiled using the Affymetrix HT U133 Plus arrays recapitulates that previously reported using Affymetrix U133A arrays[54] **c,d**, Gene set enrichment analysis demonstrates enrichment for the JAK-STAT signalling pathway in ETP ALL.

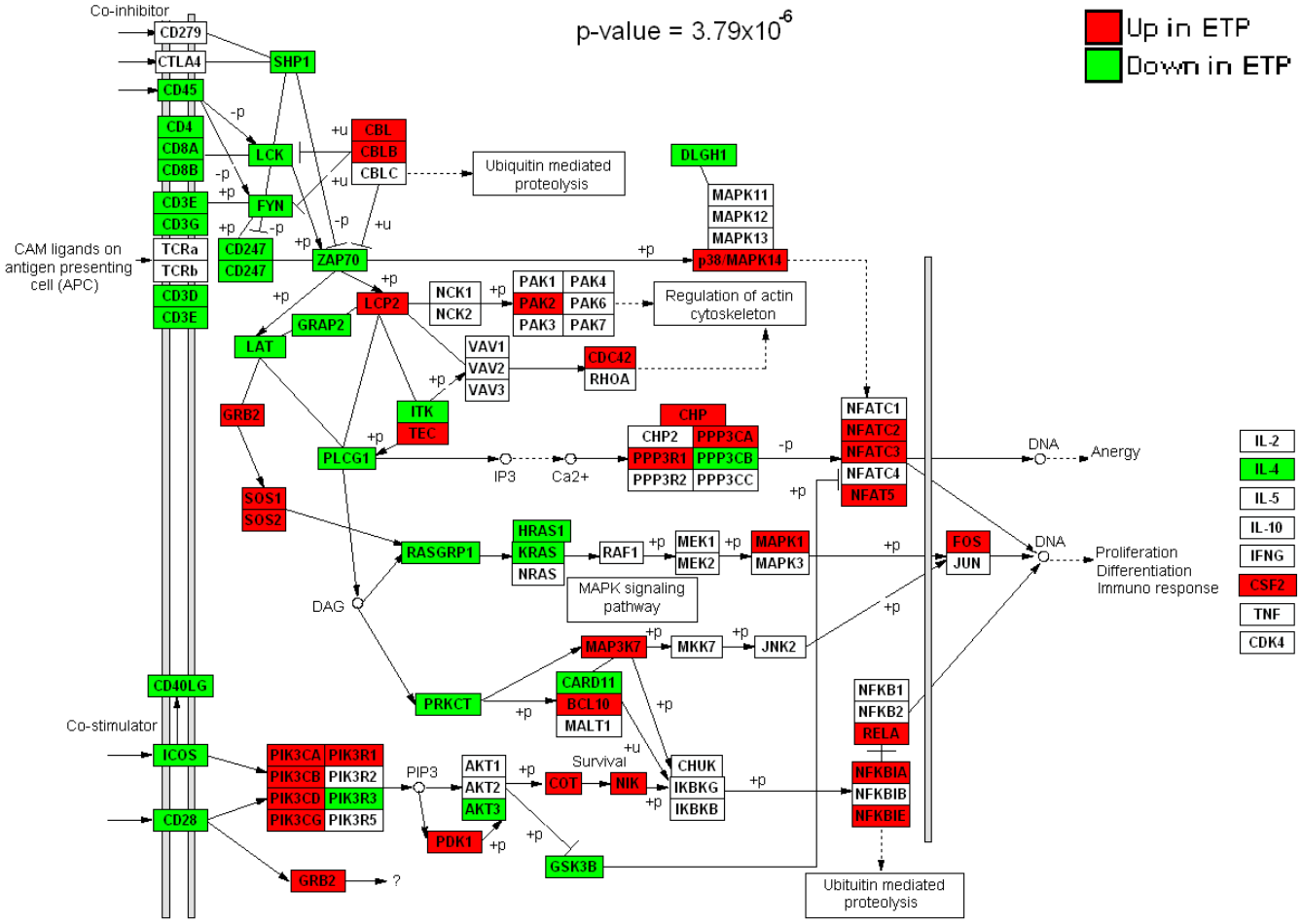
Supplementary Figure 24. Unsupervised hierarchical clustering of T-ALL showing recurring genetic alterations

The top panel shows unsupervised heirarchical clustering of minimally filtered microarray gene expression data, showing clustering of ETP ALL cases. The middle panel depicts trimmed expression of genes defining T-ALL subtypes. The lower panel shows recurring alterations in genes (Green) and pathways (red) in ETP and non-ETP T-ALL. Cases in grey lack mutation recurrence data.

Supplementary Figure 25. DAVID analyses of dysregulated pathway expression in ETP ALL

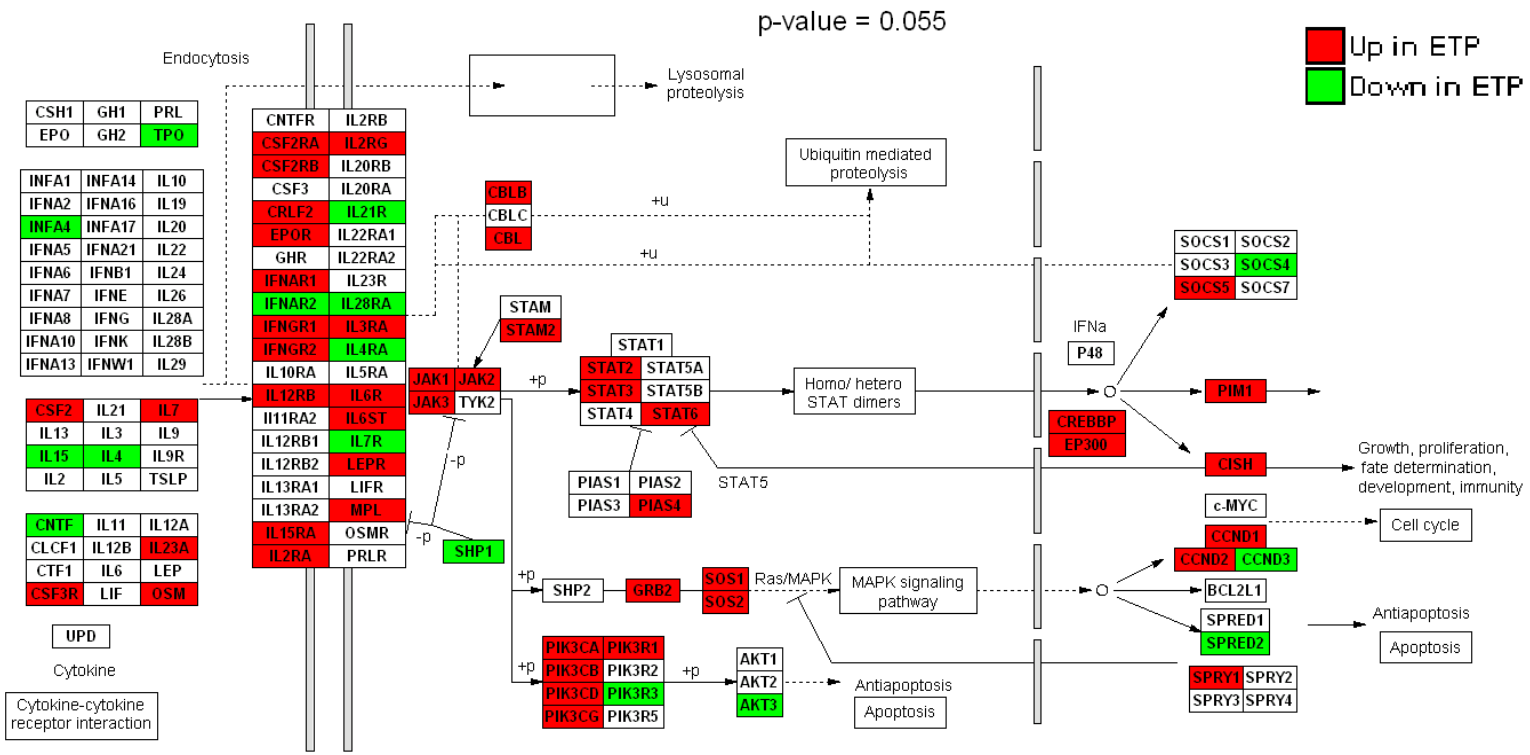
DAVID analysis showing **a**, reduced expression of T-cell signalling genes; and **b,** increased expression of JAK-STAT signalling genes in T-lineage ALL. Probe sets differentially expressed between ETP and non-ETP T-ALL to an FDR threshold of 10% were used as input for DAVID analyses.

a

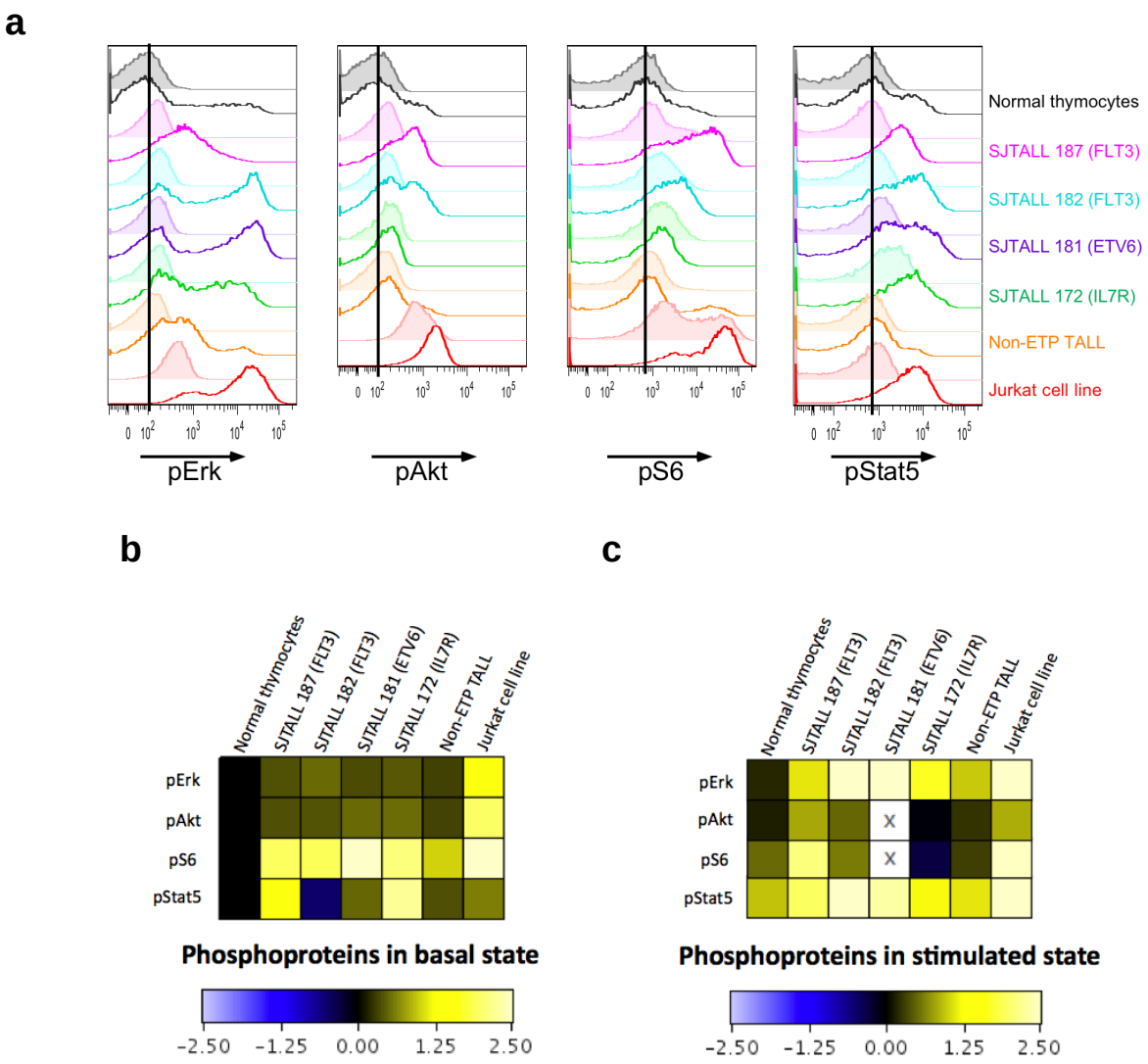


a

b

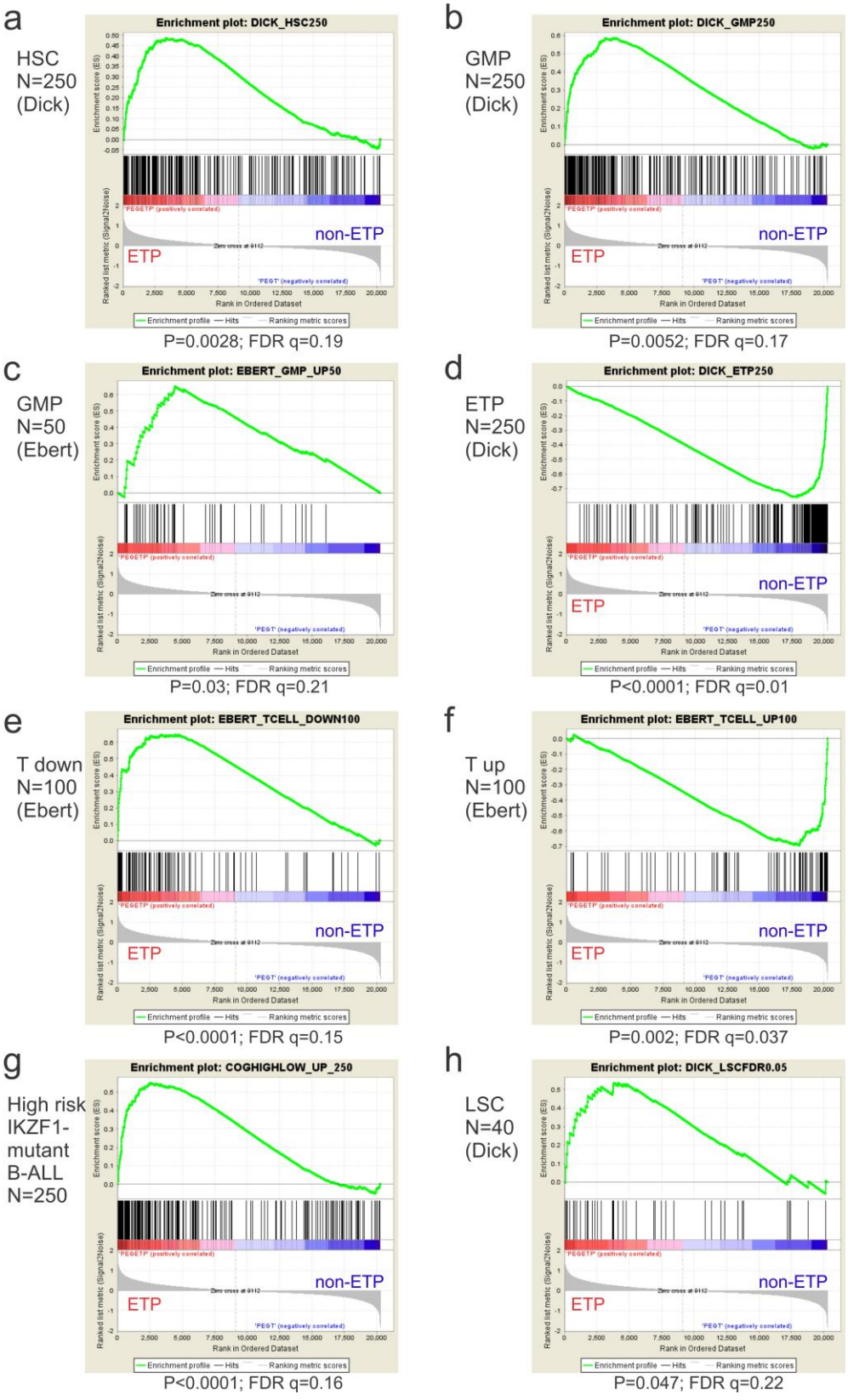

p-value = 0.055
Up in ETP
Down in ETP
Endocytosis
Lysosomal proteolysis
Ubiquitin mediated proteolysis
CSH1 GH1 PRL
EPO GH2 TPO
INFA1 INFA14 IL10
IFNA2 IFNA16 IL19
INFA4 INFA17 IL20
IFNA5 IFNA21 IL22
IFNA6 IFNB1 IL24
IFNA7 IFNE IL26
IFNA8 IFNG IL28A
IFNA10 IFNK IL28B
IFNA13 IFNW1 IL29
CSF2 IL21 IL7
IL13 IL3 IL9
IL15 IL4 IL9R
IL2 IL5 TSLP
CNTF IL11 IL12A
CLCF1 IL12B IL23A
CTF1 IL6 LEP
CSF3R LIF OSM
UPD
Cytokine
Cytokine-cytokine receptor interaction
CNTFR IL2RB
CSF2RA IL2RG
CSF2RB IL20RB
CSF3 IL20RA
CRLF2 IL21R
EPOR IL22RA1
GHR IL22RA2
IFNAR1 IL23R
IFNAR2 IL28RA
IFNGR1 IL3RA
IFNGR2 IL4RA
IL10RA IL5RA
IL12RB IL6R
Il11RA2 IL6ST
IL12RB1 IL7R
IL12RB2 LEPR
IL13RA1 LIFR
IL13RA2 MPL
IL15RA OSMR
IL2RA PRLR
CBLB
CBLC
CBL
+u
+u
STAM
STAM2
JAK1 JAK2
JAK3 TYK2
+p
-p
-p
SHP1
STAT1
STAT2 STAT5A
STAT3 STAT5B
STAT4 STAT6
Homo/ hetero STAT dimers
PIAS1 PIAS2
PIAS3 PIAS4
STAT5
IFNa
P48
CREBBP
EP300
SOCS1 SOCS2
SOCS3 SOCS4
SOCS5 SOCS7
PIM1
CISH
Growth, proliferation, fate determination, development, immunity
c-MYC
Cell cycle
CCND1
CCND2 CCND3
BCL2L1
SPRED1
SPRED2
Antiapoptosis
Apoptosis
SPRY1 SPRY2
SPRY3 SPRY4
+p
SHP2
GRB2
SOS1
SOS2
Ras/MAPK
MAPK signaling pathway
+p
PIK3CA PIK3R1
PIK3CB PIK3R2
PIK3CD PIK3R3
PIK3CG PIK3R5
+p
AKT1
AKT2
AKT3
Antiapoptosis
Apoptosis

Supplementary Figure 26. Phosphosignalling analysis of ETP and non-ETP T-ALL



Phosphoflow cytometry analysis of representative primary ETP patient samples of the indicated genotype was performed. Normal human thymocytes, a non-ETP T-ALL sample, and the T-ALL line Jurkat are included as controls. **a,b** Signalling profiles in response to stimulation of MAPK, PI3K, and IL7 pathways reveals distinct signatures in representative ETP subsets relative to normal human thymocytes. **a** Histograms depicting the activation of phospho-ERK, phospho-AKT, phospho-S6, and phospho-Stat5 in the basal state and in response to stimulation with pervanadate. **b** A heat map of the median phosphorylation state in the basal state for the indicated epitopes. The intensity gradient represents the transformed ratio of medians of leukaemia sample basal state relative to normal human thymocyte control basal state for each sample. **c** A heatmap of the median phosphorylation state in the stimulated state for the indicated epitopes. The intensity gradient represents the transformed ratio of medians of the stimulated state relative to the basal state for each sample in response to pervanadate.

Supplementary Figure 27. Gene set enrichment analysis of ETP ALL



SUPPLEMENTARY REFERENCES

1 Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).

2 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).

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

4 Mullighan, C. G. *et al.* Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. *Nat Genet* **41**, 1243-1246 (2009).

5 Mullighan, C. G. *et al.* BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**, 110-114 (2008).

6 Kitagawa, Y. *et al.* Prevalent involvement of illegitimate V(D)J recombination in chromosome 9p21 deletions in lymphoid leukemia. *J Biol Chem* **277**, 46289-46297 (2002).

7 Fugmann, S. D., Lee, A. I., Shockett, P. E., Villey, I. J. & Schatz, D. G. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu Rev Immunol* **18**, 495-527 (2000).

8 Cowell, L. G., Davila, M., Ramsden, D. & Kelsoe, G. Computational tools for understanding sequence variability in recombination signals. *Immunological reviews* **200**, 57-69 (2004).

9 Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).

10 Morrison, A. J. & Shen, X. Chromatin remodelling beyond transcription: the INO80 and SWR1 complexes. *Nat Rev Mol Cell Biol* **10**, 373-384 (2009).

11 Wang, L. C. *et al.* The TEL/ETV6 gene is required specifically for hematopoiesis in the bone marrow. *Genes & development* **12**, 2392-2402 (1998).

12 Shurtleff, S. A. *et al.* TEL/AML1 fusion resulting from a cryptic t(12;21) is the most common genetic lesion in pediatric ALL and defines a subgroup of patients with an excellent prognosis. *Leukemia* **9**, 1985-1989 (1995).

13 Barjesteh van Waalwijk van Doorn-Khosrovani, S. *et al.* Somatic heterozygous mutations in ETV6 (TEL) and frequent absence of ETV6 protein in acute myeloid leukemia. *Oncogene* **24**, 4129-4137 (2005).

14 Bohlander, S. K. ETV6: a versatile player in leukemogenesis. *Semin Cancer Biol* **15**, 162-174 (2005).

15 Graux, C. *et al.* Fusion of NUP214 to ABL1 on amplified episomes in T-cell acute lymphoblastic leukemia. *Nat Genet* **36**, 1084-1089 (2004).

16 Ralston, S. H., Langston, A. L. & Reid, I. R. Pathogenesis and management of Paget's disease of bone. *Lancet* **372**, 155-163 (2008).

17 Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-1806 (2008).

18 Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-1812 (2008).

19 Tao, Y., Pinzi, V., Bourhis, J. & Deutsch, E. Mechanisms of disease: signaling of the insulin-like growth factor 1 receptor pathway--therapeutic perspectives in cancer. *Nature clinical practice. Oncology* **4**, 591-602 (2007).

20 Gregory, S. G. *et al.* Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat Genet* **39**, 1083-1091 (2007).

21 Trevino, L. R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1001-1005 (2009).

22 Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1006-1010 (2009).
23 Campbell, P. J. Somatic and germline genetics at the JAK2 locus. *Nat Genet* **41**, 385-386 (2009).
24 Mullighan, C. G. *et al.* CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature* **471**, 235-239 (2011).
25 Pasqualucci, L. *et al.* Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature* **471**, 189-195 (2011).
26 Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298-303 (2011).
27 Kasper, L. H. *et al.* A transcription-factor-binding surface of coactivator p300 is required for haematopoiesis. *Nature* **419**, 738-743 (2002).
28 Fukuyama, T. *et al.* Histone acetyltransferase CBP is vital to demarcate conventional and innate CD8+ T-cell development. *Mol Cell Biol* **29**, 3894-3904 (2009).
29 Kasper, L. H. *et al.* Conditional knockout mice reveal distinct functions for the global transcriptional coactivators CBP and p300 in T-cell development. *Mol Cell Biol* **26**, 789-809 (2006).
30 Wang, L. C. *et al.* Yolk sac angiogenic defect and intra-embryonic apoptosis in mice lacking the Ets-related factor TEL. *EMBO J* **16**, 4374-4383 (1997).
31 Georgopoulos, K. *et al.* The Ikaros gene is required for the development of all lymphoid lineages. *Cell* **79**, 143-156 (1994).
32 Georgopoulos, K., Moore, D. D. & Derfler, B. Ikaros, an early lymphoid-specific transcription factor and a putative mediator for T cell commitment. *Science* **258**, 808-812 (1992).
33 Winandy, S., Wu, P. & Georgopoulos, K. A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. *Cell* **83**, 289-299 (1995).
34 Mullighan, C. G. *et al.* Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. *N Engl J Med* **360**, 470-480 (2009).
35 Den Boer, M. L. *et al.* A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol* **10**, 125-134 (2009).
36 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).
37 Mullighan, C. G. *et al.* Next Generation Transcriptomic Resequencing Identifies Novel Genetic Alterations in High-Risk (HR) Childhood Acute Lymphoblastic Leukemia (ALL): A Report From the Children's Oncology Group (COG) HR ALL TARGET Project. *Blood* **114**, abstract 704 (2009).
38 Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758-764 (2007).
39 Okuda, T., van Deursen, J., Hiebert, S. W., Grosveld, G. & Downing, J. R. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84**, 321-330 (1996).
40 Lorsbach, R. B. *et al.* Role of RUNX1 in adult hematopoiesis: analysis of RUNX1-IRES-GFP knock-in mice reveals differential lineage expression. *Blood* **103**, 2522-2529 (2004).
41 Egawa, T., Tillman, R. E., Naoe, Y., Taniuchi, I. & Littman, D. R. The role of the Runx transcription factors in thymocyte differentiation and in homeostasis of naive T cells. *The Journal of experimental medicine* **204**, 1945-1957 (2007).
42 Wong, W. F., Kohu, K., Chiba, T., Sato, T. & Satake, M. Interplay of transcription factors in T-cell differentiation and function: the role of Runx. *Immunology* **132**, 157-164 (2011).

43 Downing, J. R. *et al.* An AML1/ETO fusion transcript is consistently detected by RNA-based polymerase chain reaction in acute myelogenous leukemia containing the (8;21)(q22;q22) translocation. *Blood* **81**, 2860-2865 (1993).
44 Dicker, F. *et al.* Mutation analysis for RUNX1, MLL-PTD, FLT3-ITD, NPM1 and NRAS in 269 patients with MDS or secondary AML. *Leukemia* **24**, 1528-1532 (2010).
45 Gaidzik, V. I. *et al.* RUNX1 Mutations in Acute Myeloid Leukemia: Results From a Comprehensive Genetic and Clinical Analysis From the AML Study Group. *J Clin Oncol* (2011).
46 Gelsi-Boyer, V. *et al.* Genome profiling of chronic myelomonocytic leukemia: frequent alterations of RAS and RUNX1 genes. *BMC Cancer* **8**, 299 (2008).
47 Song, W. J. *et al.* Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. *Nat Genet* **23**, 166-175 (1999).
48 Churpek, J. E. *et al.* Identification and molecular characterization of a novel 3' mutation in RUNX1 in a family with familial platelet disorder. *Leuk Lymphoma* **51**, 1931-1935 (2010).
49 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).
50 Makishima, H. *et al.* Novel homo- and hemizygous mutations in EZH2 in myeloid malignancies. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K* **24**, 1799-1804 (2010).
51 Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nature genetics* **42**, 722-726 (2010).
52 Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat Genet* **42**, 181-185 (2010).
53 Sneeringer, C. J. *et al.* Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 20980-20985 (2010).
54 Coustan-Smith, E. *et al.* Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. *Lancet Oncol* **10**, 147-156 (2009).
55 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).
56 Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-663 (2007).
57 Mullighan, C. G. Single nucleotide polymorphism microarray analysis of genetic alterations in cancer. *Methods in molecular biology* **730**, 235-258 (2011).
58 Pounds, S. *et al.* Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* **25**, 315-321 (2009).
59 Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72 (2008).
60 Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-1066 (2009).
61 Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).
62 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196 (2010).
63 Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
64 Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Research* **20**, 413-427 (2010).

65 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).
66 Marcucci, G., Haferlach, T. & Dohner, H. Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **29**, 475-486 (2011).
67 De Keersmaecker, K. *et al.* The TLX1 oncogene drives aneuploidy in T cell transformation. *Nature medicine* **16**, 1321-1327 (2010).
68 Bezrookove, V. *et al.* A novel t(6;14)(q25-q27;q32) in acute myelocytic leukemia involves the BCL11B gene. *Cancer genetics and cytogenetics* **149**, 72-76 (2004).
69 Leroy, H. *et al.* CEBPA point mutations in hematological malignancies. *Leukemia* **19**, 329-334 (2005).
70 Goodman, R. H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. *Genes Dev* **14**, 1553-1577 (2000).
71 Cave, H. *et al.* ETV6 is the target of chromosome 12p deletions in t(12;21) childhood acute lymphocytic leukemia. *Leukemia* **11**, 1459-1464 (1997).
72 Mullighan, C. G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377-1380 (2008).
73 Zhang, J. *et al.* Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* **118**, 3080-3087 (2011).
74 Silva, F. P. *et al.* ETV6 mutations and loss in AML-M0. *Leukemia* **22**, 1639-1643 (2008).
75 Taketani, T. *et al.* FLT3 mutations in the activation loop of tyrosine kinase domain are frequently found in infant ALL with MLL rearrangements and pediatric ALL with hyperdiploidy. *Blood* **103**, 1085-1088 (2004).
76 Armstrong, S. A. *et al.* FLT3 mutations in childhood acute lymphoblastic leukemia. *Blood* **103**, 3544-3546 (2004).
77 Marcais, A. *et al.* Genetic inactivation of Ikaros is a rare event in human T-ALL. *Leukemia research* **34**, 426-429 (2010).
78 Jager, R. *et al.* Deletions of the transcription factor Ikaros in myeloproliferative neoplasms. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K* **24**, 1290-1298 (2010).
79 Shochat, C. *et al.* Gain-of-function mutations in interleukin-7 receptor-alpha (IL7R) in childhood acute lymphoblastic leukemias. *J Exp Med* **208**, 901-908 (2011).
80 Flex, E. *et al.* Somatically acquired JAK1 mutations in adult acute lymphoblastic leukemia. *J Exp Med* **205**, 751-758 (2008).
81 Bercovich, D. *et al.* Mutations of JAK2 in acute lymphoblastic leukaemias associated with Down's syndrome. *Lancet* **372**, 1484-1492 (2008).
82 Xiang, Z. *et al.* Identification of somatic JAK1 mutations in patients with acute myeloid leukemia. *Blood* **111**, 4809-4812 (2008).
83 Walters, D. K. *et al.* Activating alleles of JAK3 in acute megakaryoblastic leukemia. *Cancer Cell* **10**, 65-75 (2006).
84 Tefferi, A. Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms: JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K* **24**, 1128-1138 (2010).
85 Molteni, C. G. *et al.* PTPN11 mutations in childhood acute lymphoblastic leukemia occur as a secondary event associated with high hyperdiploidy. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K* **24**, 232-235 (2010).
86 Paulsson, K. *et al.* Mutations of FLT3, NRAS, KRAS, and PTPN11 are frequent and possibly mutually exclusive in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **47**, 26-33 (2008).

87 Yamamoto, T. *et al.* PTPN11, RAS and FLT3 mutations in childhood acute lymphoblastic leukemia. *Leukemia research* **30**, 1085-1089 (2006).
88 Tartaglia, M. *et al.* Genetic evidence for lineage-related and differentiation stage-related contribution of somatic PTPN11 mutations to leukemogenesis in childhood acute leukemia. *Blood* **104**, 307-313 (2004).
89 Loh, M. L. *et al.* Mutations in PTPN11 implicate the SHP-2 phosphatase in leukemogenesis. *Blood* **103**, 2325-2331 (2004).
90 Nomdedeu, J. *et al.* Low frequency of exon 3 PTPN11 mutations in adult de novo acute myeloid leukemia. Analysis of a consecutive series of 173 patients. *Haematologica* **90**, 412-413 (2005).
91 Nishimoto, N. *et al.* T cell acute lymphoblastic leukemia arising from familial platelet disorder. *International journal of hematology* **92**, 194-197 (2010).
92 Rand, V. *et al.* Genomic characterization implicates iAMP21 as a likely primary genetic event in childhood B-cell precursor acute lymphoblastic leukemia. *Blood* (2011).
93 Tang, J. L. *et al.* AML1/RUNX1 mutations in 470 adult patients with de novo acute myeloid leukemia: prognostic implication and interaction with other gene alterations. *Blood* **114**, 5352-5361 (2009).
94 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).
95 Cobb, B. S. *et al.* Targeting of Ikaros to pericentromeric heterochromatin by direct DNA binding. *Genes Dev* **14**, 2146-2160 (2000).
96 Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343-349 (2011).