Supplementary data
Supplementary Methods

Gene-expression analyses

Raw counts from RNA-seq data were calculated using STAR 2.5.2b\(^1\) with the parameter --quantMode GeneCounts. Genes with no count in any sample were discarded. Reads per million mapped reads (RPM) were log2(x+1) transformed and quantile normalized (normalize.quantiles R function).

Expression analysis was performed using weighted correlation network analysis (WGCNA) Bioconductor R routine\(^2\) and unsupervised bi-clustering. For WGCNA soft power was set equal to 12. WGCNA was primarily used to identify and omit genes which expression is mostly related to non-leukemic marrow and blood cells within leukemic samples. WGCNA identified several modules of genes highly correlated, including the brown and the black modules as candidates for containing a large number of genes co-expressed by blood or bone marrow normal cell populations. Analysis of 5 non-leukemic samples, three from blood and two from bone marrow confirmed high expression in those samples. Moreover, expression was inversely related to the percentage of leukemic cells in samples used for molecular analyses. In addition, a large proportion of those genes was annotated by non-redundant genesets from subpopulations of the cell atlas.\(^3\) We showed that maintaining those genes in the list of genes subjected to differential analyses (volcano plot and hypergeometric analysis) had dramatic consequence on interpretation, therefore those genes were omitted. Similar approach was done for the module corresponding to genes expressed by the Y chromosome in male patients. The pink module was identified as highly correlated in cases from the CDX2/UBTF cluster. The lists of genes in these modules are provided in Supplementary Tables 2-5. Other modules were more complex and were not further annotated. For unsupervised bi-clustering, the dChip software was used (http://www.dchip.org/). For data visualization, the Complex Heatmap R routine (Bioconductor, ComplexHeatmap) and the t-distributed stochastic neighbor embedding (tSNE) analysis, R package Rtsne with a perplexity value of 30 was used. Comparison between expression means was analyzed by the moderate T-test (Welch test). Enrichments against gene lists were evaluated by the GSEA method (https://www.gsea-msigdb.orgGSEA, GSEA 4.2.1, version).

RNA-seq mutation detection

RNA-seq data were also used for mutation detection. SAMtools 1.2\(^4\) was used to generate mpileup and variant calling was performed using VarScan2 mpileup2cns 2.4.0.\(^5\) The variants were annotated with Annovar\(^6\) and only exonic non-synonymous variants in 660 cancer genes were considered. High-probability oncogenic mutations were retained by eliminating sequencing/mapping errors and known/possible SNPs based on the available databases.
Targeted next-generation sequencing (NGS)

Targeted NGS was performed on 276 ALL cases in order to identify mutations and copy-number aberrations on a custom panel of 189 known or putative target genes in B-ALL (listed in Supplementary Table 13), using custom capture-based target enrichment (SureSelect, Agilent, Santa Clara, CA, USA) and sequencing on Illumina NextSeq500 platform (Illumina San Diego, CA, USA), as previously described. Capture-based target enrichment and sequencing of a 333 kb region in the 13q12.2 locus (chr13:27,962,069-28,295,338; hg38) was also used to determine deletion breakpoints in B-ALL cases having 13q12.2 deletions. BAM files were examined using the Integrative Genomics Viewer (IGV).

Array-CGH analysis

Array-CGH analysis was performed on 131 ALL cases using Agilent 400K oligonucleotide arrays, following the manufacturer’s recommendations. The arrays were scanned using the SureScan High-Resolution Technology (Agilent Technologies). Data analysis was performed using Agilent Cytogenomics software.

Sequential RNA-DNA FISH

Experiments were performed as previously described. Briefly, PDX cells were spotted on glass slides in presence of ribonucleoside-vanadyl complex (New England Biolabs) and incubated with the CDX2 fluorescent probes in a dark and humid chamber at 37°C overnight (RNA FISH part). Nuclei were counterstained with DAPI. After image acquisition with recording of field coordinates, mounting medium was wash off, slides were incubated in RNaseA (New England Biolabs), post-permeabilized, then denatured before overnight incubation with the Fosmid fluorescent probes in a dark and humid chamber at 42°C (DNA FISH part). Slides were then washed, counterstained, mounted as previously and recorded fields of view were imaged again.

Probe preparation. Probes for CDX2 primary transcripts (hgCDX2 cloned in three part in pSUPER.PKD1.RNAi), Fosmid G248P8885B5 and Fosmid G248P8671B10 were directly labeled by nick translation (Vysis kit, Abbott molecular) using Aminoallyl-dUTP-ATTO-550, 488 and 647, respectively (Jena Bioscience), following manufacturer’s instructions. Around 0.3µg of labeled products was ethanol precipitated with 10µg of Salmon sperm (Thermofisher). 5µg of human Cot-1 DNA (Thermofisher) were also added with the Fosmid probes for subsequent pre-annealing of potential repeated sequences. Pellets were resuspended in 10µl of sterile formamide. Probes were then denatured and either incubated for pre-annealing (Fosmid probes), or kept on ice until hybridization with cells (CDX2 probe). 10µg of 2X hybridization medium (4X SSC, 20% dextran sulfate, 4mg/ml BSA (New England Biolabs), 2mM ribonucleoside-vanadyl complex (New England Biolabs) were added to the probes just prior to overnight hybridization.

Microscopy. 3D-image stacks, with optical sections separated by 0.24µm, were recorded on a IXplore spinning disk microscope (Olympus) with an Orca-flash-4.0 V3 CMOS camera (Hamamatsu) and a 60X
objective. After RNA FISH, coordinates of fields of view were recorded in order to be imaged back after the DNA FISH part. Stacks were analyzed using Image I software.

**Patient-derived xenografts (PDX) of CDX2/UBTF ALL**

PDX of CDX2/UBTF ALL were established using NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NOD-SCID gamma-null, or NSG) mice\(^9\) according to experimental procedures previously described.\(^10\) Mice were housed and handled in the pathogen-free animal facility Département d’Expérimentation Animale in accordance with the guidelines of the Animal Care and Use Committee (IRSL, Saint-Louis Hospital). Authorization for animal experimentation was obtained, APAFIS#19382-2019022115238547v3, in accordance with French laws. Primary leukemia cells were CD3 depleted from cryopreserved patients’ diagnostic samples and injected intrabone into NSG mice previously irradiated with 1.25 Gy. Engraftment was monitored by flow cytometry analysis of bone marrow aspirates using human PE-CD19 and Pacific Blue-CD45 antibodies (Miltenyi). Mice were euthanized when they reached endpoints set to meet accepted animal care guidelines, and spleen and bone marrow cells were harvested. The stability of CDX2/UBTF ALL-specific transcriptome and genome features upon xenotransplantation was controlled by RNA-seq and capture 13q12.2 NGS on 2 and 1 cases, respectively. Human leukemic cells from PDX were used for RNA-DNA FISH and ChIP-seq experiments.

**Chromatin immunoprecipitation sequencing (ChIP-seq)**

ChIP assays were performed using iDeal ChIP-seq kit for Transcription Factors (Diagenode) following the manufacturer’s recommendations. Briefly, 3 million cells from PDX or cell lines were incubated for 10 min in 1% formaldehyde in phosphate buffered saline at room temperature, quenched by adding 1/10 volume of glycine. Chromatin was sonicated in 1.5 ml Bioruptor\(^\circledast\) Pico Microtubes using a Bioruptor Pico instrument for 20 cycles of sonication [30sec ON, 30sec OFF]. Immunoprecipitations were performed with H3K27ac antibody (C15410196), H3K4me3 antibody (C15410003) and a normal rabbit IgG control (Diagenode). ChIP-seq libraries were prepared using MicroPlex Library Preparation Kit v3 (Diagenode) according to manufacturer’s instructions. Libraries underwent 75-cycle single-end sequencing on a Nextseq500 (Illumina, San Diego, CA, USA). Sequence reads were mapped to reference genome hg38 using Bowtie 1.0.0 with the following parameters -m 1 --strata --best -y -S -I 40 -p 2. Bigwig files were generated using Homer\(^\circledast\) software makeUCSCfile script with default parameters and scaled to 1e7 reads.

**ATAC-seq**

ATAC-seq was performed on cell lines using a protocol adapted from Buenrostro et al.\(^{12}\) Briefly, 50000 cells were lysed and nuclei were treated with Nextera Tn5 transposase using the Nextera DNA library kit. DNA was purified and amplified using NEBNext High-Fidelity enzyme (NEB) and primers described in Buenrostro et al.\(^{13}\) Purified libraries were sequenced on an Illumina HiSeq 4000 sequencer as paired-end 100 bases reads by the GenomEast platform. Image analysis and base calling were performed using RTA 2.7.3 and
bcl2fastq 2.17.1.14. Data analysis was performed using the Encode ATA-seq pipeline v1.9.3. Adapter sequences were removed and low-quality ends were trimmed. Sequence alignment was performed into the hg38 assembly of Homo Sapiens genome using Bowtie2 (version 2.2.6)\(^\text{14}\) choosing the zero multi-mapping option. Mitochondrial reads were removed. Fold-enrichment bigwig files were generated with the MACS2\(^\text{15}\) bdgcmp command.

**4C-seq**

4C-seq was performed following the protocol described in Karasu et al.\(^\text{16}\) Briefly, 10 million cells were fixed then lysed. Nuclei were permeabilized and the chromatin was digested using DpnII enzyme (NEB), circularized using T4 DNA ligase (NEB), digested using Csp6I enzyme (NEB) and recircularized. Inverse PCR was performed using the ExpandTM Long Template PCR System (Roche) and Cdx2 specific primers containing Illumina indexes and adapter sequences: Csp6I primer matches to position hg38 chr13:27,969,087-27,969,104 and DpnII primer to position hg38 chr13:27,969,870-27,969,889 around the transcription start site of **CDX2** gene.

PCR reactions were pooled, primers removed by washing with 1.8x AMPure XP beads, then quantified on a Bioanalyzer (Agilent) before sequencing. Libraries were sequenced on an Illumina HiSeq 4000 sequencer as paired-end 100 bases reads by the GenomEast platform, a member of the ‘France Genomique’ consortium (ANR-10-INBS-0009). Image analysis and base calling were performed using RTA version 2.7.7 and bcl2fastq version 2.20.0.422. All bait sequence (including and downstream of the primer sequence, up to but not including the GATC DpnII site) are trimmed by the demultiplexing Sabre tool (https://github.com/najoshi/sabre), allowing two mismatches, before mapping to hg38 human genome assembly with Bowtie.\(^\text{17}\) Intrachromosomal reads were assigned to DpnII fragments by utility tools coming with the 4See package,\(^\text{18}\) which was also used to visualize the 4C profiles. The interaction with d3 enhancer was called with peakC\(^\text{19}\) robust to various window sizes.

**CRISPR/Cas9 interference**

The catalytically inactive Cas9 (dead Cas9) fused to the repressor KRAB was used to inactivate the **PAN3** enhancer. Single guide sequences targeted the **PAN3** enhancer were designed using E-CRISpR,\(^\text{20}\) sgEnhancer#1, GCATTTCAGAATGTTTATAG; sgEnhancer#5, GCATTACAATAGGGGTAGAG. sgRNA were inserted into the dCas9-KRAB-T2a-GFP lentiviral backbone using annealed oligos and the BsmBI cloning site of pLV hU6-sgRNA hUbC-dCas9-KRAB-T2a-GFP (plasmid #71237, AddGene).

Lentiviruses were produced with jetPRIME reagents (Polyplus transfection) by co-transfection of viral production plasmids in T150 cm\(^2\) tissue culture flasks containing 2 million HEK-293T cells by flask according to manufacturer’s protocol. As previously described\(^\text{21}\), HEK-293T cells were incubated at 37°C for 18 hours and the medium was replaced with fresh complete Advanced DMEM medium. Viral supernatant was harvested 48 hours later and filtered through a 0.45um HV Durapore membrane (Merck Millipore) and
concentrated by ultra-centrifugation at 24,000 rpm for 2 hours at 4°C. Viral supernatant was resuspended in 1X phosphate-buffered saline (PBS) and stored at -80°C.

NALM16 cells were transduced overnight for 16 hours with lentiviral supernatants of sgEnhancer#1, sgEnhancer#5 or a control with a gRNA sequence not present in the human genome, sgRenilla (GGAACACGGCCGTATTAGGG). Cells were FACS sorted on GFP-expression 72 hours after infection.

**RT-qPCR**

CDX2 expression level was measured by RT-qPCR performed on the StepOne (Thermo Fisher) using the TaqMan Multiplex Master Mix (Thermo Fisher). Each sample was analyzed in duplicate, and gene expression was normalized relative to the expression of the housekeeping gene ABL1. Primers (forward, F and reverse, R) and probes (P) sequences were: CDX2 F-GCGGAACCTGTGCGAGTG, R-GACTGTAGTGAACTCTTCTCCAGC and P-CGCGCAGCAGTCCCTCG, ABL1 F-TGGAGATAACACTAAGCATAACTAAGGT, R-GATGATTTGCTTTGGGACCCA and P-CCATTTTTGTTGGCTTCACACATT.

Primers and probe for UBTF::ATXN7L3 fusion transcript were: F- AGGAGCAGCAAAGACGTACA, R-GACCAGGTCCCGTGATATCTC and P-AGTCAGAGCGCGTGCTCCACAAGAGTTT.

**Western blot**

Cells were washed with cold PBS and lysed for 30min with RIPA buffer containing protease and phosphatase inhibitor cocktail (Thermo Scientific). Lysates were centrifuged at 25,000 g for 30 min at 4°C. Protein concentration was quantified with QuantiPro BCA Assay Kit (Sigma). 5µg of the protein extracts were denatured, separated on 4%–15% Mini-PROTEAN TGX gels (BioRad) for 80min at 90-130V and electrotransferred onto 0.2µm PVDF membranes for 30min in a Trans-Blot Turbo equipment (BioRad). Membranes were blocked with 3% fat-free BSA in TBS-T buffer (Tris buffered saline buffer pH 7.5 and 0.1% Tween 20) for 60 min and incubated overnight at 4°C with primary antibodies (CDX2, #12306, Cell Signaling (dilution 1:1000), ATXN7L3, A302-800A, Bethyl (dilution 1:2000), GAPDH, G8795, Sigma (dilution 1:1000) or vinculin antibody, ab130007, Abcam (dilution 1:1000)). After rinsing with TBS-T, membranes were incubated with an HRP-conjugated anti-rabbit or mouse IgG secondary antibody for 60 min. Immunoblots were visualized with SuperSignal™ West Femto Maximum Sensitivity Substrate (Thermo Fisher) and imaged in a Fusion FX equipment (Vilber). Band density for CDX2 protein were normalized against vinculin protein content and analysis performed by the ImageJ software.

**Statistical analysis**

Comparisons for categorical and continues variables between CDX2/UBTF ALL and other B-ALL were performed with Fisher’s exact test and Mann–Whitney test, respectively. Overall survival (OS) was calculated from the date of diagnosis to the last follow-up date censoring patients alive. The cumulative
incidence of relapse (CIR) was calculated from the complete remission date to the date of relapse, censoring patients alive without relapse at the last follow-up date. Relapse and death in complete remission were considered as competitive events. Univariate and multivariate analyses assessing the impact of categorical and continuous variables were performed with a Cox model. Proportional-hazards assumption was checked before conducting multivariate analyses. The multivariate analysis include age, log(WBC), post-induction MRD (≥10^-4 versus < 10^-4), and CDX2 status as covariates. Statistical analyses were performed with STATA software (STATA 16.1, StataCorp LLC, College Station, TX). All p-values were two-sided, with p < 0.05 denoting statistical significance.
Supplementary Tables

Supplementary Table 1. Comparison between GRAALL-2005/2014 studied and non-studied patients.

|                      | Studied patients (n=723) | Non-studied patients (n=291) | P value |
|----------------------|--------------------------|-------------------------------|---------|
| **Patient-related characteristics** |                          |                               |         |
| Median age, years [range] | 38 [18-60]               | 39 [18-60]                    | 0.62    |
| Gender, M/F (ratio)      | 400/323 (1.2)            | 157/134 (1.2)                 | 0.73    |
| **Disease-related characteristics** |                          |                               |         |
| Median WBC, G/L [range]  | 9 [0-712]                | 4 [0-294]                     | <0.001  |
| BM blast percentage [range] | 91 [21-100]             | 84 [20-98]                    | 0.003   |
| CNS involvement         | 57/713 (8%)              | 13/290 (4%)                   | 0.055   |
| **Response-related characteristics** |                          |                               |         |
| Good PB prednisone response at day 8 | 582/701 (83%)            | 253/286 (88%)                 | 0.033   |
| Good BM response at day 15 | 397/668 (59%)            | 182/278 (65%)                 | 0.092   |
| Late CR (achieved after induction 2) | 24/672 (4%)              | 8/266 (3%)                    | 0.84    |
| CR (after induction 1 and 2) | 672/723 (93%)            | 266/291 (91%)                 | 0.43    |
| Post-induction MRD ≥ 10^-4 | 255/542 (47%)            | 43/119 (36%)                  | 0.033   |
| Post-induction MRD ≥ 10^-3 | 151/542 (28%)            | 33/119 (28%)                  | 1.0     |
| **Post-remission outcome** |                          |                               |         |
| Allo-SCT in first CR     | 184/672 (27%)            | 83/266 (31%)                  | 0.26    |
| 3-years CIR, % (95% CI)  | 33.5 (29.8-37.6)         | 18.7 (14.4-24.1)              | <0.001  |
| 3-years OS, % (95% CI)   | 65.9 (62.0-69.5)         | 69.6 (63.7-74.7)              | 0.45    |

M, male; F, female; WBC, white blood cell count; BM, bone marrow; CNS, central nervous system; PB, peripheral blood; CR, complete remission; MRD, minimal residual disease. Allo-SCT, allogeneic stem cell transplantation; CIR, cumulative incidence of relapse; OS, overall survival. Good PB prednisone response at day 8 is defined by less than 1G/L blast. Good BM response at day 15 is defined by less than 5% blast. MRD evaluation was performed by PCR quantification of Ig/TR clono-specific rearrangements according to EuroMRD guidelines.
Supplementary Table 8. Breakpoint coordinates of the 13q12.2 deletions.

| Patient Id | Deleted region (hg38) | Oncogenic subtype | Deletion involving FLT3 |
|------------|------------------------|-------------------|-------------------------|
| B_SL7      | chr13:28098868-28163018 | CDX2/UBTF | Yes |
| B_SL19     | chr13:28098540-28161424 | CDX2/UBTF | Yes |
| B_SL39     | chr13:28098526-28149074 | CDX2/UBTF | Yes |
| B_SL43     | chr13:28098507-28246295 | CDX2/UBTF | Yes |
| B_SL49     | chr13:28098535-28149074 | CDX2/UBTF | Yes |
| B_SL73     | chr13:28098864-28158208 | CDX2/UBTF | Yes |
| B_SL79     | chr13:28098533-28208033 | CDX2/UBTF | Yes |
| B_SL82     | chr13:28098523-28186795 | CDX2/UBTF | Yes |
| B_SL100    | chr13:28098536-28140159 | CDX2/UBTF | Yes |
| B_SL139    | chr13:28098867-28181482 | CDX2/UBTF | Yes |
| B_SL141    | chr13:28098535-28181482 | CDX2/UBTF | Yes |
| B_SL157    | chr13:28098860-28140158 | CDX2/UBTF | Yes |
| B_SL160    | chr13:28098529-28208035 | CDX2/UBTF | Yes |
| B_SL168    | chr13:28098540-28181483 | CDX2/UBTF | Yes |
| B_SL187    | chr13:28031725-28141296 | CDX2/UBTF | Yes |
| B_SL220    | chr13:28098868-28141296 | CDX2/UBTF | Yes |
| B_SL234    | chr13:28098537-28149079 | CDX2/UBTF | Yes |
| B_SL251    | chr13:28098542-28166955 | CDX2/UBTF | Yes |
| B_SL271    | chr13:28098867-28150357 | CDX2/UBTF | Yes |
| B_SL283    | chr13:28098531-28181486 | CDX2/UBTF | Yes |
| B_SL327    | chr13:28098868-28143877 | CDX2/UBTF | Yes |
| B_SL366    | chr13:28098544-28149074 | CDX2/UBTF | Yes |
| B_SL425    | chr13:28098538-28181483 | CDX2/UBTF | Yes |
| CDX2_SL1   | chr13:28098537-28181482 | CDX2/UBTF | Yes |
| CDX2_SL2   | chr13:28098531-28181486 | CDX2/UBTF | Yes |
| CDX2_SL3   | chr13:28098541-28149077 | CDX2/UBTF | Yes |
| B_SL57     | chr13:28101793-28242690 | Other | No |
| B_SL78     | chr13:28101756-28243897 | Other | No |
| B_SL184    | chr13:28101792-28242694 | Other | No |
| B_SL186    | chr13:28101756-28166951 | Other | No |
| B_SL188    | chr13:28101791-28242689 | Other | No |
| B_SL205    | chr13:28101795-28242691 | Other | No |
| B_SL214    | chr13:28101763-28166943 | ZNF384 fusion | No |
| B_SL259    | chr13:28098857-28246271 | ZNF384 fusion | Yes |
| B_SL287    | chr13:28129155-28242695 | Other | No |
| B_SL287    | chr13:28114035-28240996 | Other | No |
| B_SL313    | chr13:28105013-28151976 | Other | No |
Supplementary Table 9. CDX2 allelic expression data showing monoallelic expression in all informative CDX2/UTBF cases.

| Patient Id | gDNA rs1805108 | gDNA rs1805107 | mRNA rs1805108 | mRNA rs1805107 | CDX2 expression |
|------------|----------------|----------------|----------------|----------------|----------------|
| B_SL7      | C/A            | A/G            | A              | G              | Monoallelic    |
| B_SL19     | C/C            | A/A            | Non informative|               |                |
| B_SL39     | C/A            | A/G            | A              | G              | Monoallelic    |
| B_SL43     | C/C            | A/A            | Non informative|               |                |
| B_SL49     | C/C            | A/A            | Non informative|               |                |
| B_SL73     | C/A            | A/G            | A              | G              | Monoallelic    |
| B_SL79     | C/A            | A/G            | A              | G              | Monoallelic    |
| B_SL82     | C/C            | A/A            | Non informative|               |                |
| B_SL100    | C/C            | A/A            | Non informative|               |                |
| B_SL139    | C/C            | A/A            | Non informative|               |                |
| B_SL141    | C/C            | A/A            | Non informative|               |                |
| B_SL157    | C/A            | A/G            | C              | A              | Monoallelic    |
| B_SL160    | C/A            | A/G            | A              | G              | Monoallelic    |
| B_SL168    | C/C            | A/A            | Non informative|               |                |
| B_SL187    | C/C            | A/A            | Non informative|               |                |
| B_SL220    | C/C            | A/A            | Non informative|               |                |
| B_SL234    | C/A            | A/A            | Non informative|               |                |
| B_SL251    | C/A            | A/G            | A              | G              | Monoallelic    |
| B_SL271    | A/A            | G/G            | Non informative|               |                |
| B_SL283    | C/A            | A/G            | C              | A              | Monoallelic    |
| B_SL287    | C/A            | A/G            | C              | A              | Monoallelic    |
| B_SL366    | A/A            | G/G            | Non informative|               |                |
| B_SL425    | C/C            | A/A            | Non informative|               |                |
| CDX2_SL1   | A/A            | G/G            | Non informative|               |                |
| CDX2_SL2   | C/A            | A/G            | NA             |                |                |
| CDX2_SL3   | C/A            | A/G            | A              | G              | Monoallelic    |

Supplementary Table 10. RNA-DNA FISH data showing that CDX2 is expressed from the chromosome 13 exhibiting the deletion.

| Per nucleus: | N   | %   |
|--------------|-----|-----|
| B_SL157      |     |     |
| CDX2 expressed from the WT chromosome | 0   | 0.0 |
| CDX2 expressed from the deleted chromosome | 78  | 61.4|
| B_SL141      |     |     |
| CDX2 expressed from the WT chromosome | 0   | 0.0 |
| CDX2 expressed from the deleted chromosome | 18  | 15.1|

- **CDX2 RNA part**
- **G248P8671810 DNA part**
- **G248P888585 DNA part**

B_SL157: 61.4% of the cells showed CDX2 expression (always from the deleted chromosome).
B_SL141: 15.1% of the cells show CDX2 expression (always from the deleted chromosome).
Supplementary Table 12. Cytogenetic data of the 26 patients with CDX2/UBTF ALL

| Patient Id | Karyotype | 1q gain (technique) |
|------------|-----------|---------------------|
| B_SL7      | 46,XX,del(1)(p31?q33),del(2)(p?13p22),del(9)(q1?3?q34),?del(12)(p1?3),?del(17)(p?1?),+mar,inc[18]/46,idem,?X[2] | No |
| B_SL19     | 46,XY[32] | No |
| B_SL39     | 46,XY,del(1)[q41],[t(2;6)[q23?q1?2],?del(6)[6qter?6q23?1q2?3?4?6qter][21] | No |
| B_SL43     | 46,XX,del(3)[q2?5],inc[cp8]/46,XX[4] | No |
| B_SL49     | 46,XX,del(15)[t(1;15)[p12;p11][15]/46,XX,del(11)[t(?1;11)[p12;p15][3]/46,XX[4] | Yes (karyo, CGH, NGS) |
| B_SL73     | 46,XX,del(9)[13q31][24]/46,XX[3] | No |
| B_SL79     | Failure | No |
| B_SL82     | 47,XX,?+8,der(1;15)[q10;p10][12]/46,XX[8] | Yes (karyo, CGH, NGS) |
| B_SL100    | 46,XX,del(12)[p12][7]/46,sl,t(11;18)[q22?q12][3]/46,del(9)[q2?2?4][5]/46,XX[5] | No |
| B_SL141    | 46,XX,del(12)[p12][7]/46,sl,t(11;18)[q22?q12][3]/46,del(9)[q2?2?4][5]/46,XX[5] | No |
| B_SL157    | 46,XX[20] | No |
| B_SL160    | 46,XX,del(7)[p13][3]/46,XX[8] | No |
| B_SL168    | 46,XX,del(1)[q12?q32][6]/46,XX[4] | Yes (karyo) |
| B_SL187    | 46,XY,15ps+[20] | Yes (CGH, NGS) |
| B_SL220    | 46,XX,del(1)[q2?2?4][3]/46,del(4)[q2?2?4][2]/46,XX[22] | Yes (karyo, CGH, NGS) |
| B_SL234    | 46,XY[20] | No |
| B_SL251    | 46,XX,del(3)[p22],add(7)[q21],add(12)[q24][5]/46,XX[15] | No |
| B_SL271    | 81?95,XXYY,del(1)[q2?2?4][q2?2?4][x2,add(2)[p173][7][3][4],?del(6)[q2?q25][x2?9,16,?20,?5,?3?5?12]/46,XY[1] | Yes (karyo, CGH, NGS) |
| B_SL283    | 46,X,Y,?3,?15,?18,?5?mar[5]/46,XY[10] | Yes (CGH, NGS) |
| B_SL327    | 46,XX,del(7)[7;13][p171?q21][3]/46,sl,add(3)[q2?7?6],del(9)[q13?31] or [q2?2?3][2]/46,XX[7] | Yes (CGH, NGS) |
| B_SL366    | 46,XX,del(18)[1;18][q25?q23][25] | Yes (karyo, CGH, NGS) |
| B_SL425    | 46,XX,add(2)[q37],del(9)[q13?q33]<17>/46,XX<3] | Yes (NGS) |
| CDX2_SL1   | 92<4n>,XX,?X,?add(7)[x28],?8,del(9)[q13?q22][x2,add(18)[q22][x2,add(2)[p21]/46,XX[4] | No |
| CDX2_SL2   | 46,XX,del(4)[t(1;4)[q11?q37][11]/92<4n>,XX,del(9)[q22][4]/46,XX[18] | Yes (karyo, CGH, NGS) |
| CDX2_SL3   | 46,XX,inc[3]/42?46,XX,inc,?6?11?mar[3]/46,XY[38] | No |

karyo, karyotype; CGH, array comparative hybridization; NGS, next-generation sequencing.
Supplementary Table 14. Full list of mutations identified in CDX2/UBTF ALL.

| Patient id | Gene | Mutation | VAF (%) |
|------------|------|----------|---------|
| B_SL7      | CXCR4| NM_003467:exon2:c.1013C>G:p.(Ser338Ter) | 46.1 |
| B_SL19     | U2AF1| NM_001025203:exon2:c.101C>T:p.(Ser34Phe) | 22.5 |
| B_SL39     | No mutation | | |
| B_SL43     | KRAS | NM_033360:exon4:c.351A>T:p.(Lys117Asn) | 29.2 |
| B_SL43     | DNMT3A| NM_022552:exon16:c.1913C>G:p.(Arg736His) | 31.5 |
| B_SL49     | CREBBP| NM_004380:exon16:3090_3091insATGA:p.(Gln1031MetfsTer3) | 31.5 |
| B_SL49     | TP53 | NM_000546:exon8:c.782A>G:p.(Lys261Asp) | 5.0 |
| B_SL49     | CXCR4| NM_003467:exon2:1030_1031insCGGGCCAC:p.(Ser344SerfsTer25) | 17.7 |
| B_SL49     | CXCR4| NM_003467:exon2:1000C>T:p.(Arg334Ter) | 13.1 |
| B_SL49     | KRAS | NM_033360:exon4:c.351A>T:p.(Lys117Asn) | 29.2 |
| B_SL49     | DNMT3A| NM_022552:exon16:c.1913C>G:p.(Arg736His) | 31.5 |
| B_SL73     | No mutation | | |
| B_SL43     | NRAS | NM_002524:exon2:c.35G>A:p.(Gly12Asp) | 4.6 |
| B_SL77     | No mutation | | |
| B_SL82     | DNMT3A| NM_175629:exon14:c.1578C>A:p.(Tyr526Ter) | 3.4 |
| B_SL82     | CREBBP| NM_004380:exon31:c.6361_6396del:p.(Leu2121_Gly2132del) | 9.8 |
| B_SL100    | TET2 | NM_001127208:exon3:c.2725_2734del:p.(Gln909ArgfsTer9) | 35.7 |
| B_SL100    | CXCR4| NM_003467:exon2:1000C>T:p.(Arg334Ter) | 10.4 |
| B_SL139    | PAX5 | NM_016734:exon6:c.676_677insAGGGGG:p.(Arg225_Gly226insGluGly) | 6.8 |
| B_SL141    | No mutation | | |
| B_SL157    | No mutation | | |
| B_SL160    | NRAS | NM_002524:exon2:c.35G>C:p.(Gly12Ala) | 2.6 |
| B_SL160    | MEF2C| NM_001193347:exon12:c.1319C>T:p.(Ser440Leu) | 21.0 |
| B_SL168    | No mutation | | |
| B_SL187    | CREBBP| NM_004380:exon26:c.4336C>T:p.(Arg1446Cys) | 19.4 |
| B_SL187    | CXCR4| NM_003467:exon2:c.952_953insGGG:p.(Thr318ArgfsTer15) | 11.6 |
| B_SL220    | No mutation | | |
| B_SL234    | NRAS | NM_002524:exon2:c.38G>A:p.(Gly13Asp) | 35.7 |
| B_SL251    | CREBBP| NM_004380:exon3:c.901C>T:p.(Gln301Ter) | 39.8 |
| B_SL271    | No mutation | | |
| B_SL283    | ERG | NM_001243432:exon9:c.801G>A:p.(Trp267Ter) | 15.4 |
| B_SL283    | FLT3 | NM_004119:exon20:c.2503G>T:p.(Asp835Tyr) | 29.5 |
| B_SL327    | No mutation | | |
| B_SL366    | NRAS | NM_002524:exon2:c.35G>C:p.(Gly12Ala) | 3.9 |
| B_SL366    | KRAS | NM_033360:exon2:c.35G>A:p.(Gly12Asp) | 9.7 |
| B_SL425    | ARID1A| NM_006015:exon20:c.6259G>A:p.(Gly2087Arg) | 36.2 |
| B_SL425    | KRAS | NM_033360:exon4:c.436G>A:p.(Ala146Thr) | 3.1 |
| B_SL425    | CXCR4| NM_003467:exon2:c.1000C>T:p.(Arg334Ter) | 13.2 |
| B_SL425    | CXCR4| NM_003467:exon2:1004_1005insC:p.(Gly335GlyfsTer9) | 16.4 |
| CDX2_SL1   | No mutation | | |
| CDX2_SL2   | KRAS | NM_033360:exon2:c.35G>C:p.(Gly12Ala) | 4.5 |
| CDX2_SL2   | PTPN11| NM_002834:exon3:c.214G>A:p.(Ala72Thr) | 8.2 |
| CDX2_SL2   | CREBBP| NM_004380:exon4:c.1022_1023insCCGC:p.(Ala341AlafsTer10) | 8.5 |
| CDX2_SL2   | KRAS | NM_033360:exon2:c.38G>A:p.(Gly13Asp) | 8.9 |
| CDX2_SL2   | RUNX1| NM_001754:exon6:c.584T>C:p.(Ile195Thr) | 30.8 |
| CDX2_SL3   | No mutation | | |

VAF, variant allele frequency on targeted NGS, except for CXCR4 mutations detected on RNA-seq data.
Supplementary Table 15. Immunophenotype of CDX2/UBTF ALL.

| Patient Id | EGIL group | CD19 | CD79a | CD10 | CD20 | CD22 | clgM | CD34 | CD38 | CD13 | CD33 |
|------------|------------|------|-------|------|------|------|------|------|------|------|------|
| B_SL7      | B-II       | pos  | pos   | pos  | neg  | pos  | neg  | pos  | NA   | neg  | neg  |
| B_SL19     | B-II       | pos  | pos   | pos (54%) | neg | pos | neg | pos | NA | neg | neg |
| B_SL39     | B-II       | pos  | pos | pos | NA | NA | NA | NA | pos | NA | NA |
| B_SL43     | B-I        | pos  | pos | neg | neg | pos | neg | neg | pos | neg | neg |
| B_SL49     | B-II       | NA   | pos | pos (42%) | neg | pos (40%) | neg | NA | NA | neg | neg |
| B_SL73     | B-I        | pos  | pos | neg | neg | pos | neg | pos | pos | neg | NA |
| B_SL79     | B-I        | pos  | pos | neg | neg | NA | NA | pos | pos | neg | neg |
| B_SL82     | B-I        | pos  | neg | neg | neg | pos | neg | pos | NA | neg | neg |
| B_SL100    | B-I        | pos  | pos | neg | neg | pos | neg | pos | pos | neg | neg |
| B_SL139    | B-I        | pos  | pos | neg | NA | pos | neg | pos | pos | neg | neg |
| B_SL141    | B-I        | pos  | pos | neg | neg | neg | pos | NA | neg | neg | neg |
| B_SL157    | B-III      | pos  | pos | pos | NA | pos | pos | pos | pos | neg | neg |
| B_SL160    | B-I        | pos  | pos | neg | neg | pos | neg | pos | pos | neg | neg |
| B_SL168    | B-I        | pos  | pos | neg | neg | pos | neg | pos | pos | neg | neg |
| B_SL187    | B-I        | pos  | NA | neg | neg | pos | NA | pos | pos | neg | neg |
| B_SL220    | B-II       | pos  | pos | pos (47%) | neg | pos | neg | pos | pos | neg | neg |
| B_SL234    | B-I        | pos  | pos | neg | neg | pos | neg | pos | pos | (20%) | NA |
| B_SL251    | B-II       | pos  | pos | pos (44%) | neg | pos | pos (20%) | NA | NA | neg | neg |
| B_SL271    | B-II       | pos  | pos | pos (32%) | neg | pos | neg | pos | pos | neg | neg |
| B_SL283    | B-I        | pos  | pos | neg | pos (20%) | neg | neg | pos | NA | neg | neg |
| B_SL327    | B-I        | pos  | pos | neg | NA | pos (20%) | NA | pos | pos | neg | neg |
| B_SL366    | B-I        | pos  | pos | neg | neg | pos | neg | pos | pos | neg | pos (50%) |
| B_SL425    | B-II       | pos  | pos | pos | pos | pos | neg | pos | NA | neg | neg |
| CDX2_SL1   | B-I        | pos  | pos | neg | neg | NA | NA | pos | NA | NA | NA |
| CDX2_SL2   | B-II       | pos  | pos | pos | neg | pos | neg | NA | NA | NA | NA |
| CDX2_SL3   | B-I        | NA   | neg | neg | pos | NA | pos | NA | pos | neg | neg |

EGIL, European Group for the Immunological Characterization of Leukemias; NA, not available. Percentages are indicated in case of partial expression.
Supplementary Figures

**Supplementary Figure 1. Flow-chart describing the patients’ cohorts.** The cohort of B-ALL analyzed by RNA-seq corresponded to unresolved B-ALL with available material from patients prospectively enrolled in the GRAALL-2005/2014 and EWALL-INO protocols. Additional B-other ALL cases analyzed by RNA-seq or targeted molecular analyses (CDX2 qRT-PCR, 13q12.2 NGS and RT-PCR UBTF::ATXN7L3) for patient care and for whom a CDX2/UBTF ALL was diagnosed were included in the CDX2/UBTF ALL cohort. Patients’ characteristics and outcome of CDX2/UBTF ALL were analyzed within the GRAALL-2005/2014 studied cohort. Comparison between GRAALL-2005/2014 studied (n=723) and non-studied (n=291) patients is provided in Supplementary Table 1. *Patients with non-eligibility criteria, patients who received a different therapy, or who withdrew consent.
Supplementary Figure 2. Gene-expression analyses of RNA-seq data from the cohort of 302 patients allow identification of a novel B-ALL subtype. (A) WGCNA method was used on 6047 genes selected on the basis of variability in their expression (standard deviation >0.9). Several genes’ modules were identified, including one corresponding to genes expressed in erythroid cells (black) and one to genes expressed in myeloid and lymphoid non-leukemic cells (brown), present in patients’ samples in varying percentages, one to genes expressed by the Y chromosome in male patients, and one to genes expressed in the CDX2/UBTF ALL (pink). List of genes for these modules are provided in Supplementary Tables 2-5. (B) Hierarchical bi-clustering was performed on genes selected for based on standard deviation across samples greater than 1.9 and expression equal or greater than 7 in at least 7% of samples, and excluding genes from the WGCNA modules corresponding to genes expressed by non-leukemic cells (black and brown modules) and chromosome Y, resulting in a final list of 235 genes, provided in Supplementary Table 6. (C) Supervised hierarchical clustering using genes from the WGCNA pink module identifies the CDX2/UBTF cluster.
Supplementary Figure 3. CDX2 expression quantified by qRT-PCR in acute leukemias and control cases. CDX2/UBTF ALL patients (n=7), other B-ALL (n=38), T-ALL (n=4), AML (n=13) and normal blood or bone marrow samples (n=5). RNA CDX2 expression measured by RT-qPCR and normalized on ABL1. Comparison analysis was performed by the Wilcoxon rank-sum test.
Supplementary Figure 4. Gene-expression of *UBTF* and *ATXN7L3* in CDX2/UBTF ALL and other B-ALL based on RNA-seq counts. Gene expression was quantified by RNA-seq (log2 of counts normalized using DESeq2’s median of ratios) and comparison analysis was performed by the Wilcoxon rank-sum test.
Supplementary Figure 5. CDX2 expression quantified by qRT-PCR in CDX2/UBTF ALL and cell lines. CDX2/UBTF ALL patients (n=7), NALM16, NB4, K562 and Caco-2/TC7 cell lines. RNA CDX2 expression measured by RT-qPCR and normalized on ABL1.
Supplementary Figure 6. Breakpoint sequences of the 13q12.2 deletion including genomic mapping of RSS-like sequences. Breakpoint sequences were scanned using a position weight matrix generated from the RSS conservation table of Hesse et al22 using the software tool FIMO, part of the MEME suite. B23 A background model of 20% C/G and 30% A/T, a pseudocount of 1 and a threshold p value of 0.001 was used. Conserved sequences and n-diversity are in bold, deleted sequences are in grey, boxes highlight RSS motifs and red bases, bases shared with RSS motifs.

B_SLS7

FLT3

TCATACGTCTAGTGGAAGTGAATGCATTAAT/

+TAA

TTTAAGCCGAAATGCAAGTCAACATAGCAGTCAATC

B_SL19

FLT3

CTATACGTCTAGTGGAAGTGAATGCATTAAT/

+TAA

GTCTTCAGTAATCTGTTCAAGTCAACATAGCAGTCAATC

B_SL39

FLT3

CCATACGTCTAGTGGAAGTGAATGCATTAAT/

+CTA

AGTCTTCAGTAATCTGTTCAAGTCAACATAGCAGTCAATC

B_SL43

FLT3

ATGGTTAGAGAAAAGAGTGGAACATATAAGA/

+AGCT

AAGCTGGACATAGTGGGAGGAGTCAATC

B_SL49

FLT3

CCATACGTCTAGTGGAAGTGAATGCATTAAT/

+GCTA

AGTCTTCAGTAATCTGTTCAAGTCAACATAGCAGTCAATC

B_SL73

FLT3

ATGGTTAGAGAAAAGAGTGGAACATATAAGA/

+AGCT

AAGCTGGACATAGTGGGAGGAGTCAATC

B_SL79

FLT3

TCGCTACGTCTAGTGGAAGTGAATGCATTAAT/

+CTA

TGCTGCTGGAAATCTGTTCAAGTCAACATAGCAGTCAATC

B_SL82

FLT3

ACTGGCAAATCTGCTATGGGAAGTGAATGCATTAAT/

+GAGCT

AAGCTGGACATAGTGGGAGGAGTCAATC

19
B_SL100
FLT3
TGCCATACACTGCTAATGGGAGTGGAAATTGGTG/ GTAAGACACAAAAAAAAAGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL139
FLT3
TGAGTAAAAGAAGGTGGAGAATGGAAATTGGTG/ GTAAGACACAAAAAAAAAGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL141
FLT3
CCATACACTGCTAGGAGTGGAAATTGGTG/ GTAAGACACAAAAAAAAAGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL157
FLT3
TATAATGTGAAATAGATTTGCTAATGGAAGAATGGTGAAGAGAGAGGATGGAGAAGGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL160
FLT3
GAACTGCCATACACTGCTAGTGGGAGTGGAATTGGTGTAAAGAAGGTGGACATAAAAGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL168
FLT3
CAGTCTAGGAGTGGGAGAATGGTGAAGAGAGAGGATGGAGAAGGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL187
FLT3
AAGCAGAAGGATCCACTTGTGCAAAGGCCC
GCACAGTGGGTGGGAGGACAGTGCTGGGGAACTGAAAGAAGGCCAGCGCT
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL220
FLT3
TGAGTAAAAGAAGGTGGAGAATGGAAATTGGTG/ GTAAGACACAAAAAAAAAGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3

B_SL234
FLT3
ATACACTGCTAGTGGGAGTGGAAATTGGTG/ GTAAGACACAAAAAAAAAGTTTTGGTTTGC
AGGCTGAGTCTCAAACATCTGCACTTTCTCC
PAN3
Supplementary Figure 7. Genetic mapping of all deletions at the 13q12.2 locus and gene expression. (A) Mapping of deletions at 13q12.2 locus identified by capture-based NGS. Deletions sparing FLT3 or not are represented in black and red, respectively. The arrows represent the direction of transcription. (B) Expression analysis of CDX2, URAD, FLT3, and PAN3 according to del13q12.2 status (n=131). Gene expression was quantified by RNA-seq (log2 of counts normalized using DESeq2’s median of ratios) and comparison analysis was performed by the Wilcoxon rank-sum test. (C) Dot-plot of expression data of CDX2 and FLT3 genes based on RNA-seq data according to del13q12.2 status (n=131)
Supplementary Figure 8. HOX genes expression. (A) Hierarchical clustering of the RNAseq cohort according to the expression of CDX2 and HOX genes. (B) Expression analysis of MEIS1, HOXA9 and HOXA10. Gene expression was quantified by RNA-seq (log2 of counts normalized using DESeq2’s median of ratios) and comparison analysis was performed by the Wilcoxon rank-sum test.
Supplementary Figure 9. Outcome analyses of B-ALL patients in the GRAALL2005/2014 trials according to oncogenic subtypes (A) Cumulative incidence of relapse (B) Overall survival.
Supplementary References

1. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
2. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
3. Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp Hematol*. 2018;68:51–61.
4. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
5. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568–576.
6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
7. Passet M, Boissel N, Sigaux F, et al. PAX5 P80R mutation identifies a novel subtype of B-cell precursor acute lymphoblastic leukemia with favorable outcome. *Blood*. 2019;133(3):280–284.
8. Chaumeil J, Augui S, Chow JC, Heard E. Combined immunoﬂuorescence, RNA ﬂuorescent in situ hybridization, and DNA ﬂuorescent in situ hybridization to study chromatin changes, transcriptional activity, nuclear organization, and X-chromosome inactivation. *Methods Mol Biol*. 2008;463:297–308.
9. Shultz LD, Lyons BL, Burzenski LM, et al. Human lymphoid and myeloid cell development in NOD/LtSz-scid IL2R gamma null mice engrafted with mobilized human hemopoietic stem cells. *J Immunol*. 2005;174(10):6477–6489.
10. Clappier E, Gerby B, Sigaux F, et al. Clonal selection in xenografted human T cell acute lymphoblastic leukemia recapitulates gain of malignancy at relapse. *J Exp Med*. 2011;208(4):653–61.
11. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576–589.
12. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*. 2015;109:21 29 1-21 29 9.
13. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013;10(12):1213–1218.
14. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–359.
15. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
16. Karasu N, Sexton T. 4C-Seq: Interrogating Chromatin Looping with Circular Chromosome Conformation Capture. *Methods Mol Biol*. 2021;2157:19–34.
17. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
18. Ben Zouari Y, Platania A, Molitor AM, Sexton T. 4See: A Flexible Browser to Explore 4C Data. *Front Genet*. 2019;10:1372.
19. Geeven G, Teunissen H, de Laat W, de Wit E. peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. *Nucleic Acids Res*. 2018;46(15):e91.
20. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. *Nat Methods*. 2014;11(2):122–123.
21. Gachet S, El-Chaar T, Avran D, et al. Deletion 6q Drives T-cell Leukemia Progression by Ribosome Modulation. *Cancer Discov*. 2018;8(12):1614–1631.
22. Hesse JE, Lieber MR, Mizuuchi K, Gellert M. V(D)J recombination: a functional definition of the joining signals. *Genes Dev*. 1989;3(7):1053–1061.
23. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–1018.