SUPPLEMENTAL INFORMATION

The genomic landscape of acute lymphoblastic leukemia with intrachromosomal amplification of chromosome 21

SUPPLEMENTAL METHODS

Mononuclear cells were purified from bone marrow by density gradient centrifugation and cryopreserved in liquid nitrogen at or below -150°C in preparation for DNA and RNA extraction.

Whole genome sequencing

Manual libraries were constructed with 50-2000ng of genomic DNA utilizing the Lotus Library Prep Kit (IDT Technologies) targeting 350bp inserts. Strand specific molecular indexing is a feature associated with this library method. The molecular indexes are fixed sequences that make up the first 8 bases of read 1 and read 2 insert reads. Alternatively, the TruSeq® DNA PCR-Free Library Preparation kit (Illumina) was used with 600ng input DNA. The concentration of each library was accurately determined through qPCR (Kapa Biosystems). 2x150 paired end sequence data generated ~100Gb per normal and ~200Gb per tumor sample which lead to ~30x (normal) and 60x (tumor) haploid coverage.

Whole transcriptome sequencing

WTS was performed using TruSeq library preparation and sequenced on HiSeq 2000 or 2500 platform (Illumina). All sequence reads were paired-end and sequencing was performed by using (i) total RNA and stranded RNA-sequencing (75- or 100-base pair (bp) reads); (ii) poly(A)-selected mRNA (50-, 75- or 100-bp reads). After sequencing, quality control was conducted using FastQC[1] (version 0.11.5). Sequencing reads were then mapped to the GRCh37 human genome reference by STAR[2] (version 2.4.2a). Gene annotation downloaded from Ensembl website (http://www.ensembl.org/) was used for STAR mapping and the read-count evaluation. All the samples were sequenced with RefSeq coding region covered with 30-fold coverage ≥15% (median ± standard deviation, 37.2 ± 7.5%). BAM files were sorted and analyzed with flagstat using Samtools[3] (version 1.5).

Exome sequencing

Libraries were prepared as for WGS, and 700ng aliquots used for exome capture. Five libraries were pooled at an equimolar ratio yielding a ~3.5µg library pool prior to the hybrid capture. The library pools were hybridized with the xGen Exome Research Panel v1.0 reagent (IDT Technologies) that spans 39 Mb target region (19,396 genes) of the human genome. The concentration of each captured library pool was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the NovaSeq6000 platform (Illumina). 2x150bp sequence data was generated ~50Gb per library targeting a mean depth of coverage of 500x.

Single-cell RNA sequencing (scRNA-Seq)

Frozen mononuclear cells from four samples with iAMP21-ALL (SJBALL021901, SJBALL030370, SJBALL030434 and SJBALL030871) were thawed, counted and enriched for live cells by depletion of dead cells (Miltenyi Biotec, catalogue number 130-090-101). Isolated live cells were washed three times with 1X Phosphate-Buffered Saline (PBS, calcium and magnesium free) containing 0.04% weight/volume BSA (Thermo Fisher Scientific, catalogue number AM2616) and automatedly counted by the Countess 3 Automated Cell Counter (Thermo Fisher Scientific) and manually by the Neubauer hemocytometer. From each sample, we calculated the volume of cells required to have a desired recovery target of 8,000-10,000 cells and loaded them onto Chromium Next GEM Chip K (10X Genomics; PN-2000182) with Master Mix from Chromium Next GEM Single Cell 5' Reagent Kits v2 (Dual Index) (10X Genomics; PN-1000263), Single Cell

VDJ 5' Gel Beads (10X Genomics; PN-1000264) and Partitioning Oil (10X Genomics; PN-2000190) following standard manufacturers' protocols. Briefly, nanoliter-scale gel beads-in-emulsion (GEMs) were generated, lysed and 10x barcoded, full-length cDNA from polyadenylated mRNA was produced. Next, GEMs were broken and pooled fractions were recovered. Dynabeads MyOne SILANE (10X Genomics; PN-2000048) were used to purify the first-strand cDNA from the post GEM–RT reaction mixture. Barcoded, full-length cDNA was amplified via PCR, purified by SPRIselect Reagent (Beckman Coulter, B23318) and quantified using High Sensitivity D5000 ScreenTape (5067- 5592) with High Sensitivity D5000 Reagents (5067- 5593) at the Agilent 2200 TapeStation system. Enzymatic fragmentation and size selection were used to optimize the cDNA amplicon size. To construct the final libraries, P5, P7, i5 and i7 sample indexes, and Illumina R2 sequence (read 2 primer sequence) were added via End Repair, A-tailing, Adaptor Ligation, and Sample Index PCR (10X Genomics; Library Construction Kit, 16 reactions PN-1000190 and Dual Index Kit TT Set A, 96 reactions PN-1000215). Final library quality was assessed using High Sensitivity D1000 ScreenTape (5067- 5584) with High Sensitivity D1000 Reagents (5067- 5585) at the Agilent 2200 TapeStation system. Illumina-ready dual index libraries were sequenced at the recommended depth and aiming for 50,000 reads/cell using the Illumina NovaSeq, according to manufacturer's recommendations.

Single-cell whole genome sequencing (scWGS)

Frozen bone marrow mononuclear cells from SJBALL021901 and SJBALL030072 were thawed and stained with Alexa Fluor® 700 Mouse Anti-Human CD45 (BD Biosciences, #560566) and PE Mouse Anti-Human CD3 (BD Biosciences, #561809) and single cell fluorescence-activated cell sorting (FACS) in 96-well plate for CD3-CD45dim (leukemic blasts) and CD3+CD45bright (normal T cells). Sorted single-cell were processed by PicoPLEX Gold Single Cell DNA-Seq Kit (Takara Bio USA, catalogue number R300670), according to manufacturer's instructions. Briefly, single cells were lysed, DNA was pre-amplified by linear amplification, purified by AMPure XP (Beckman Coulter, catalogue number A63880) and exponentially amplified and indexed with Illumina primers from HT Unique Dual Index Kit (1–96) (Takara Bio USA, catalogue number 634752). Libraries were purified by AMPure XP and quality was checked by Agilent D1000 ScreenTape 5067- 5582) with D1000 Reagents (5067- 5583) prior to sequencing.

Copy number determination in individual leukemia interphase nuclei by FISH

The copy number for seven different genomic loci was determined in over 300 individual leukemia cells in order to deduce the phylogenetic branching order of genetic events that had taken place during the evolution of this leukemia sample. The seven loci tested had known copy number variation in this sample: *CDKN2A* (loss of one copy), *RB1* (loss of two copies), *ETV6* (loss of one copy), *IGH* enhancer (gain of one copy of chromosome 14), *C21orf91* (loss of one copy), *HMGN1* (gain of several copies), and *PRMT2* (loss of one copy). Frozen cells from sample SJBALL021901 were used for this analysis. Thawed, unfixed cells were applied to slides by cytocentrifugation and then fixed in 4% PFA for 5 minutes, followed by 4% PFA plus 0.05% Triton X-100 for 5 minutes, and then finally slides were placed in 70% ethanol and stored at -20° C until ready for use. All probes were prepared by nick translation of either BAC or fosmid clones, using either Seebright red 580 or Seebright green 496 dUTPs. The slides were denatured in 70% formamide, 2X SSC at 80° C for 10 minute and hybridized sequentially with the 4 probe sets that were suspended in 50% formamide, 10% dextran sulfate, and 2X SSC. Probe hybridizations were performed in a series of 4 red-green pairs: 1) *RB1/CDKN2A*, 2) *ETV6/IGH*, 3) *PRMT2/C21orf91*, 4) *HMGN1*, following the schema in **Figure 5g.** Each hybridization was carried out at 37°C overnight, then slides were washed in 50% formamide 2X SSC at 37° C for 5 minutes. Slides were then mounted in Vectashield mounting medium containing DAPI counterstain and subjected to 3D imaging. After each imaging session, coverslips were carefully removed and slides were briefly rinsed in PBS before the next hybridization was performed. The resultant 4 sets of images were then analyzed together to determine the copy number of each probe in each cell of the set of cells that were imaged. This process allows for multiplexing of all 7 probes using only 2 colors. Each cell from each image was assigned an individual identity before the analysis. Signals from early hybridizations remain during subsequent hybridizations, resulting in accumulation of new red-green signals after each additional hybridization. Comparing each set of images allows for unambiguous identification of each probe signal that appears after each hybridization. The final result is the number of copies of each probe signal in each cell in the population being tested.

SNP microarray analysis

The rawcopy algorithm[4] was used to preprocess and segment SNP microarray data for genomic copy number analysis. To adjust for aneuploidy, segmentation results were then recentered so that a segment mean of zero corresponded to normal diploid segments. Segments with recentered mean < -0.2 were called regions of loss; segments with recentered mean > +0.2 were called regions of gain.

WGS and WES analysis

Paired-end WGS and WES reads were mapped to human reference genome GRCh37 by BWA[5] (version 0.7.12). Samtools[3] (version 1.5) was used to generate chromosomal coordinate–sorted and indexed bam files, which were then processed by the Picard[6] (version 1.129). MarkDuplicates module was used to mark PCR duplications. Then the reads were realigned around potential indel regions by the GATK[7] (version 3.7) IndelRealigner module. Sequencing depth and coverage were assessed based on coding regions (~34 Mb) defined by RefSeq genes. Somatic copy number alternations were detected from paired normal and tumor WGS samples using CONSERTING[8] (version 1.0). Significant copy number gains and losses were calculated using GISTIC[9] (version 2.0.23). Somatic structural variants were detected from paired normal and tumor WGS samples using three SV callers, including Delly[10] (version 0.7.6), Lumpy[11] (version 0.3.1) and Manta[12] (version 1.1.0). The SV calls passing the default quality filters of each caller were integrated using SURVIVOR[13] (version 1.0.6). The intersected call sets were manually reviewed for the supporting soft-clipped and discordant reads using the Integrated Genome Viewer (IGV).[14] Somatic variants were called from paired normal and tumor WGS samples using three established bioinformatic tools, namely Strelka[15] (version 2), MUTECT[16] (version 2), VarScan[17] (version 2.3.8). SNVs and indels called by any two callers were retained. For the merged SNVs and indels, we applied a 14-fold and 8-fold coverage cut-off for tumor and normal, separately. Significantly mutated genes were identified using MutSig2CV[18] (version 3.11) using default parameters. Multi-omics factor analysis (MOFA) was performed as described.[19] Seventy iAMP21-ALL and 179 non-iAMP21/non-hyperdiploid-ALL patients from our recent study with matched CNV and expression data were selected for this analysis.[20] Only genes on chromosome 21 were considered.

WTS analysis

To evaluate gene expression level, the read count for each annotated gene was calculated using HTSeq[21] (version 0.11.2), and gene expression level normalization and differential expression analysis were carried out using DESeq2.[22] To evaluate the digital gene expression level, a regularized log-transformed (rlog) value was calculated by DESeq2. The ComBat function within the sva R package[23] was used to correct batch effects introduced by different library preparation strategies and sequencing lengths. The R package Rtsne[24-26] was used to map samples to a two-dimensional tSNE plot with the top 1,000 most variable genes (on the basis of median absolute deviation), with the tSNE perplexity parameter was set to 30. Mutation and copy number alteration detection from RNA-sequencing were performed using the same method as described in our previous study.[27] CICERO,[28] FusionCatcher[29] (version 1.00), and STAR-Fusion[30] (version 1.1.0) were used to detect fusions, and all reported rearrangements were manually reviewed to keep reliable alterations. Internal tandem duplications detected by CICERO were also retained. The discordant reads between chromosomes 15 and 21 were extracted and manually checked using the IGV.[14] To perform gene set enrichment analysis (GSEA),[31,32] all the genes were ranked according to the fold-change and significance from differential analysis. GSEA was then performed using molecular signatures database (MSigDB)[31] (version 6.2) C2 genes and an in-house curated list of gene sets.

Genetic ancestry analysis

For each individual, the admixture fraction was estimated using the iAdmix program,[33] and allele frequencies from the 1000 Genomes Project reference populations (European, African, Native American, East Asian, and South Asian) were used as reference.[34] The overall genetic ancestral composition for each single individual was derived based on comparison of allele frequencies between patient and reference genomes.

B-ALL subtype prediction analysis

To avoid batch effects in RNA sequencing, a given patient sample was classified in an absolute manner avoiding the need for normalization toward a reference cohort. For each subtype, we first identified a set of gene pairs such that expression of gene A was greater than gene B in the subtype, but expression of gene A was less than gene B in other subtypes. To classify a patient to a specific subtype, the pairwise comparisons between two gene expression levels was done for all the gene pairs specific to the subtype. Then a majority voting was used to predict the subtype for the patient.[35]

Mutational signature profiling and molecular timing analysis

For each somatic SNV detected from a WGS sample, its trinucleotide context was determined using a custom script to obtain the 96-channel profile for each sample. COSMIC version 3.0 mutational signature in each sample was then determined using SigProfilerSingleSample (version 1.3)[36] together with the COSMIC signature set. We used MutationTimeR[37,38] to time somatic variants relative to chromosome amplifications. Each mutation is determined to be clonal [early], clonal [late], clonal [unspecified] or subclonal, based on the multiplicity state of this mutation and the copy number of the segment on which it resides.

scRNA sequencing analysis

scRNA-seq data were aligned and quantified using the Cell Ranger (v5.0.1) pipeline (http://www.10xgenomics.com) against the human genome GRCh38 (refdata-gex-GRCh38-2020-A). Cells with mitochondria content over 10% were removed. Clusters of cells were identified using Uniform Manifold Approximation and Projection (UMAP) reduction in Seurat (4.0.6, https://satijalab.org/seurat) and characterized based on gene expression of major haemopoietic cell types. Copy number variations were inferred from single cell gene expression using inferCNV with default parameters (v.1.11.1, https://github.com/broadinstitute/inferCNV). Differential expression analysis between different clusters of cells were performed using FindMarkers function from Seurat package.

scWGS analysis

scWGS data were mapped to human reference genome GRCh37 by BWA[5] (version 0.7.12). Samtools[3] (version 1.5) was used to estimate the sequencing depth for each segment inferred from bulk WGS data of SJBALL021901 and SJBALL030072. To generate the coverage normalization index, segment coverages of individual normal cell were firstly divided by the median segment coverage across the genome, and then the median value of each genomic segment among all the normal cells was used as the normalization index. To estimate the copy number for individual tumor cells, segment coverages were first divided by the coverage normalization index, then by the median segment coverage across the genome, and multiplied by two due to diploidy.

SUPPLEMENTAL TABLES

Supplemental Table 1. Summary of the iAMP21-ALL cohort.

Supplemental Table 2. Inferred genetic ancestry.

Supplemental Table 3. Gene fusions detected from WTS.

Supplemental Table 4. *P2RY8::CRLF2* clonality.

Supplemental Table 5. B-ALL subtype prediction.

Supplemental Table 6. Differential expression analysis between iAMP21-ALL and other B-ALL.

Supplemental Table 7. Significantly enriched gene sets from GSEA.

Supplemental Table 8. Protein-coding genes on chromosome 21.

Supplemental Table 9. Differential expression analysis between rob(15;21)c and typical cases.

Supplemental Table 10. Driver genes identified in each sample.

Supplemental Table 11. MutSig2CV and GISTIC analysis.

Supplemental Table 12. Summary of mutation signature analysis.

Supplemental Table 13. Association analysis of UV signature to somatic mutations.

Supplemental Table 14. Marker genes used to define cell types in scRNA-seq.

Supplemental Table 15. Differential expression analysis between different copy number clusters.

Supplemental Table 16. Median sequencing depth for each cell from scWGS-seq.

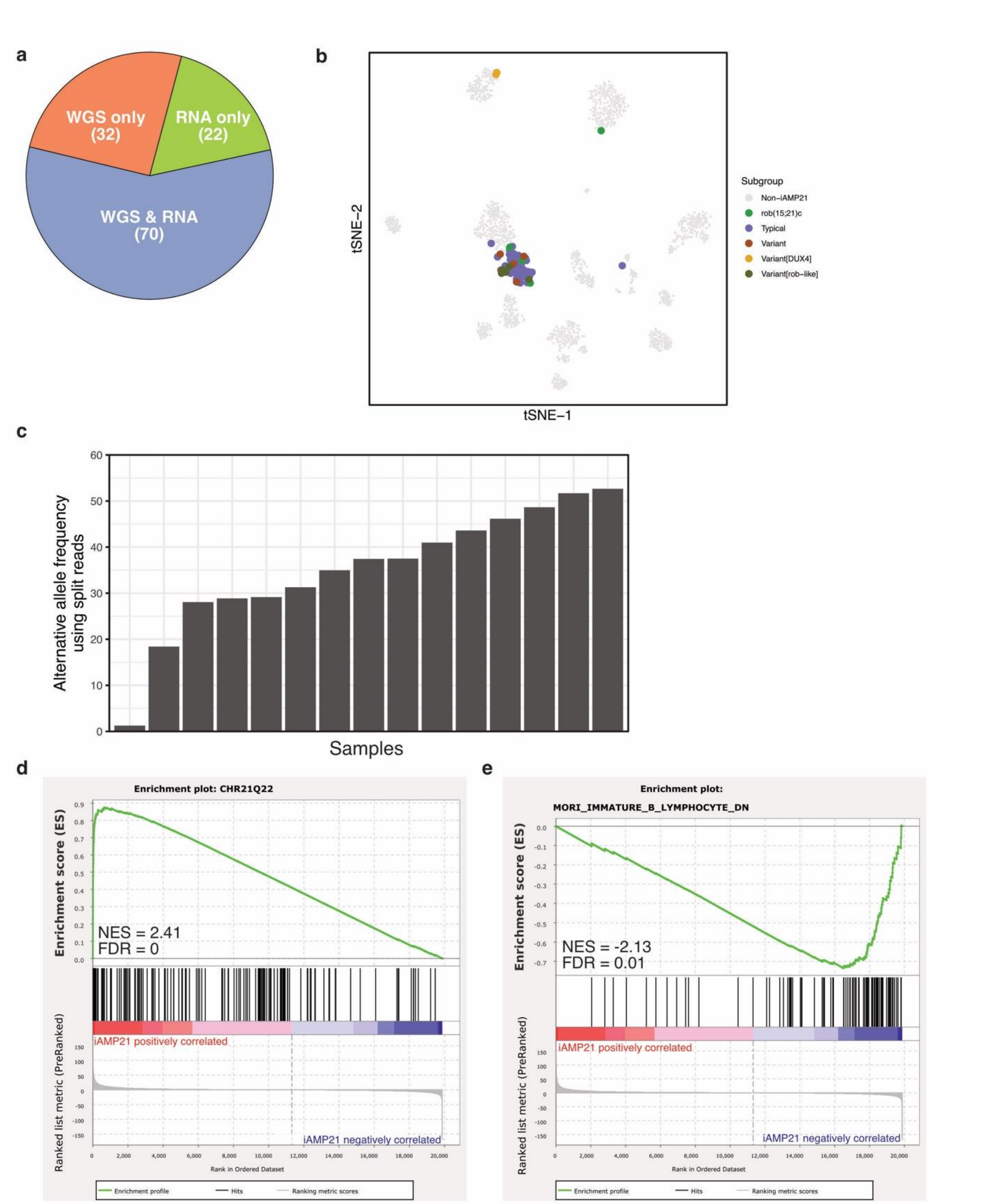
Supplemental Table 17. Segmentations for SJBALL021901 from bulk WGS.

Supplemental Table 18. Copy number and cluster for each segment in each cell from scWGS-seq.

Supplemental Table 19. Probes for FISH.

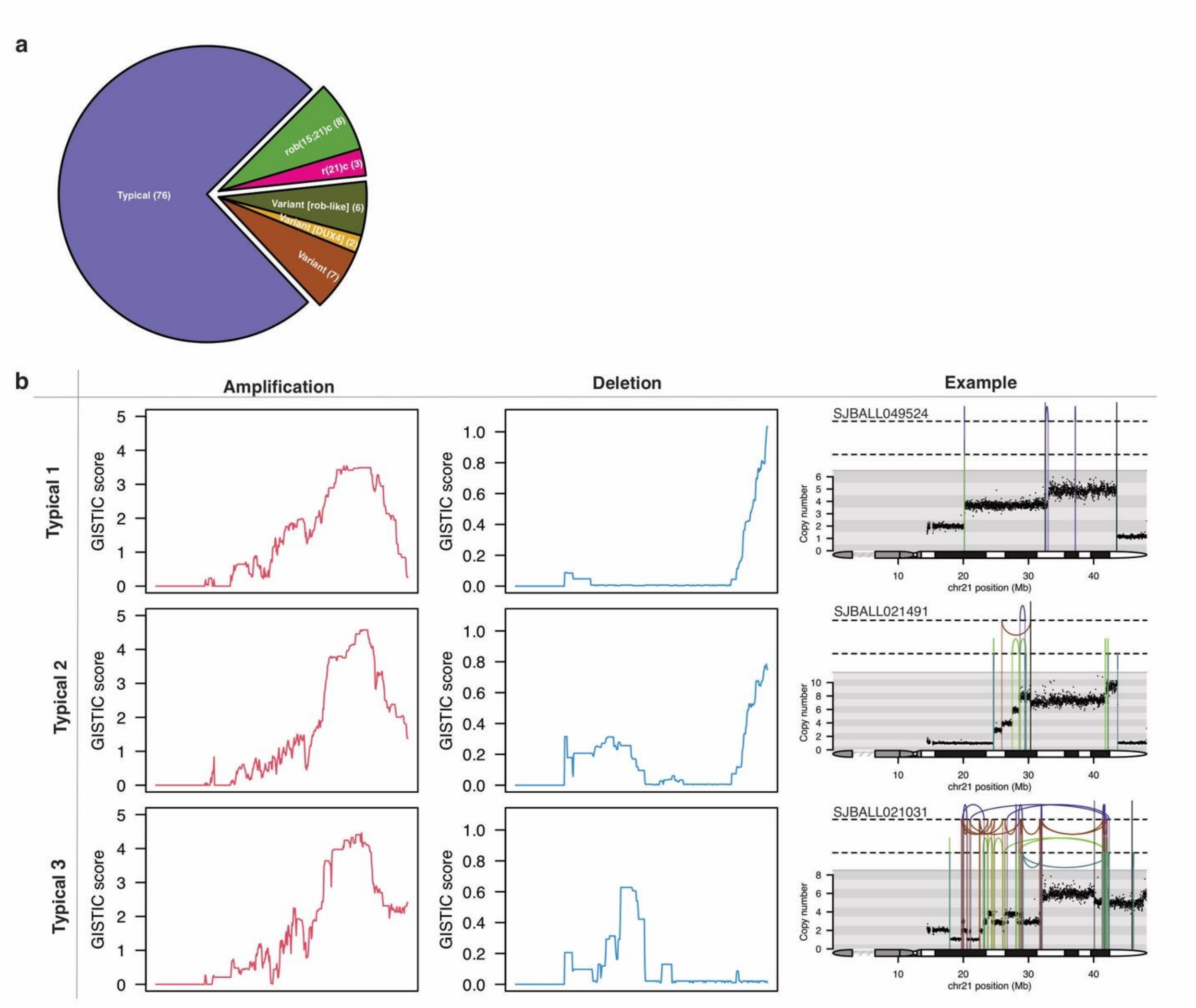
Supplemental Table 20. Copy number for *IGH*, *C21orf91*, *HMGN1* and *PRMT2* loci by FISH.

SUPPLEMENTAL FIGURES



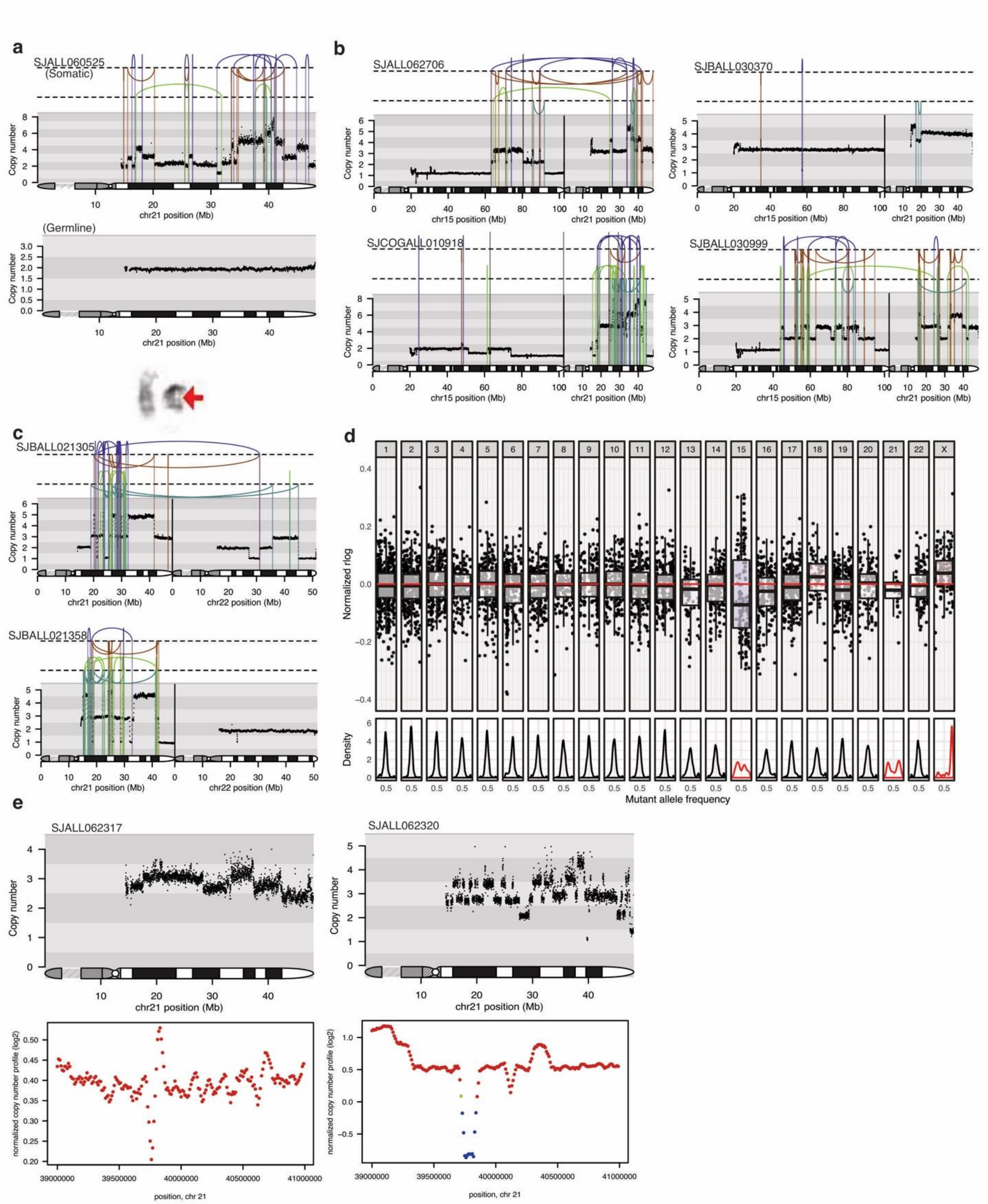
Supplemental Figure 1. Sample summary and gene expression patterns.

(a) Pie chart showing the sample distribution with respect to sequencing techniques performed. Whole genome (WGS) and whole transcriptome sequencing (WTS) data were available for 102 and 92 patients, respectively, with both datasets available for 70 cases. **(b)** t-distributed stochastic neighbor embedding (tSNE) plot showing gene expression profiling of 1,493 B-ALL samples, including 92 iAMP21-ALL. Each dot represents a sample. The top 1,000 most variable genes (on the basis of median absolute deviation) were selected and processed by the tSNE algorithm with a perplexity score of 30. Samples are color-coded based on the harbored fusion transcript. **(c)** Barplot showing alternative allele frequency using split reads (compared to total read count) that mapped to the PAR1 deletion for samples with *P2RY8::CRLF2* fusion and available WGS data. **(d)** Gene set enrichment analysis (GSEA) of iAMP21-ALL versus other B-ALL subtypes identified the chromosomal 21 region, 21q22 (CHR21Q22) as a significantly, positively enriched gene set. **(e)** Gene set enrichment analysis (GSEA) of iAMP21-ALL versus other B-ALL subtypes identified MORI_IMMATURE_B_LYMPHOCYTE_DN as a significantly, negatively enriched gene set.



Supplemental Figure 2. Case distribution and copy number profiles of typical iAMP21-ALL cases.

(a) Pie chart showing distribution of the 102 samples with WGS into iAMP21-ALL subgroups: typical (sporadic[39]) (N=76), rob(15;21)c (N=8), r(21)c (N=3), variant (N=13) and a variant subgroup associated with *DUX4*-rearrangements (N=2). **(b)** GISTIC amplification score (**left**), deletion score (**middle**) and one representative example (**right**) for the three typical subgroups.



Supplemental Figure 3. Copy number and gene expression profiles of selected cases.

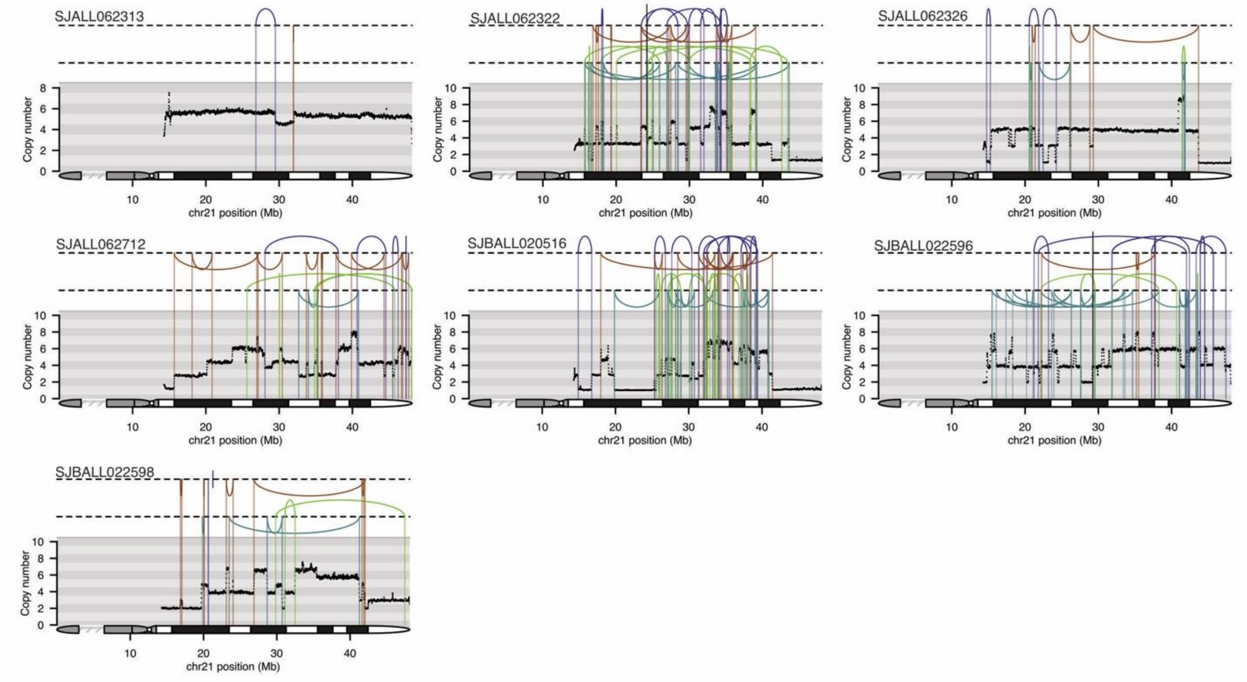
(a) Rearrangement and copy number patterns of a third case with r(21)c - SJALL060525. Both somatic and germline profiles are shown. Rearrangements are separated based on their orientation. The partial karyogram below provides evidence of r(21)c, correlating with the phenotype of the patient.

(b) Copy number and rearrangement profiles are shown of four rob(15;21)-like iAMP21-ALL patients. Copy number is shown in the y-axis and the chromosome 15 and 21 ideograms are shown along the bottom of the plot, indicating breakpoint location. Rearrangements are separated based on their orientation.

(c) Copy number and rearrangement profiles are shown for two rob(21;22)-like iAMP21-ALL patients. Copy number is shown in the y-axis and the chromosome 21 and 22 ideograms are shown along the bottom of the plot, indicating breakpoint location.

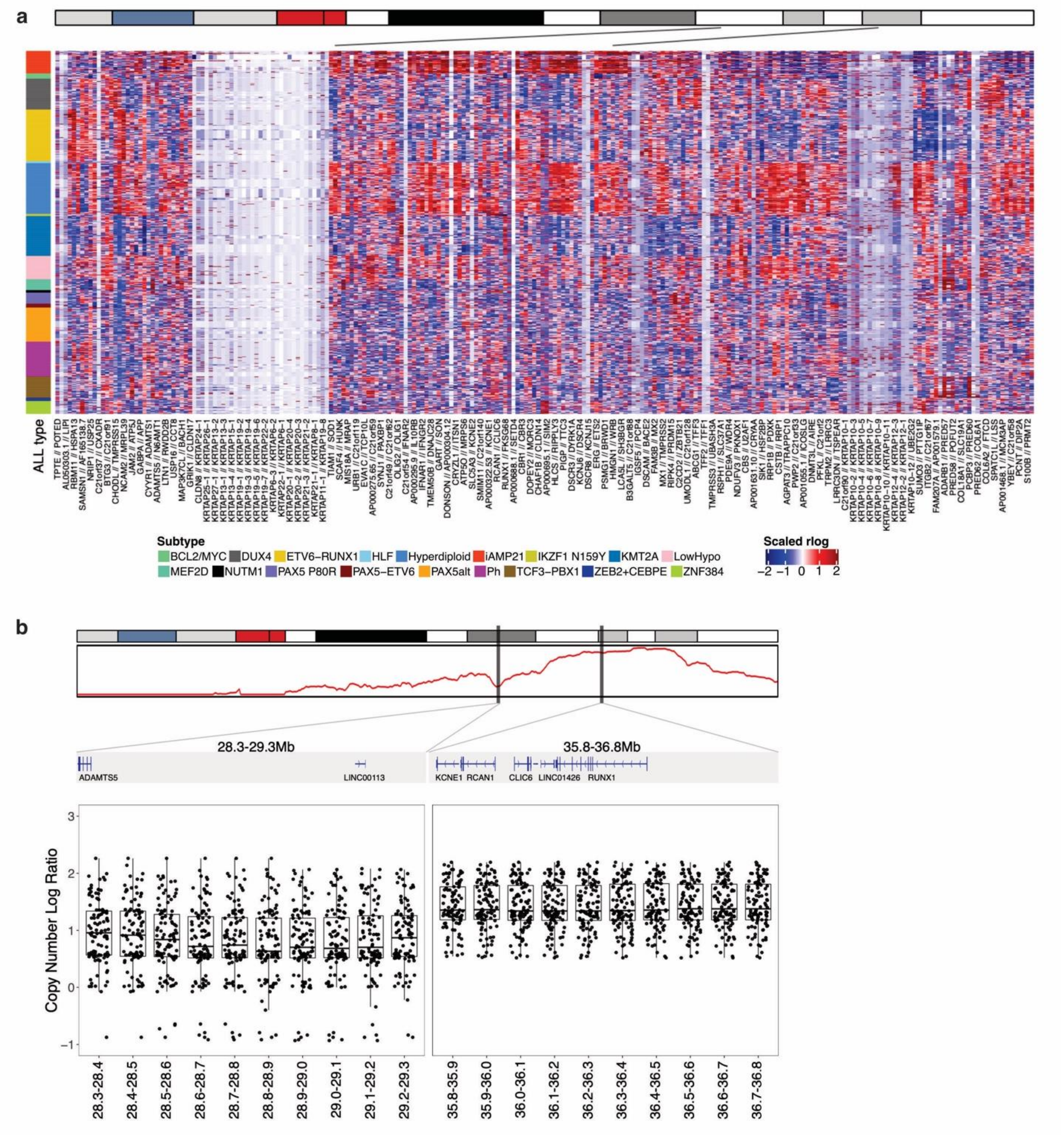
(d) Copy number derived from WTS for SJALL015979 defines this case as rob(15;21)-like as abnormal copy number profiles of both chromosomes 15 and 21 are present.

(e) Example copy number and rearrangement profiles of chromosome 21 are shown for two cases with iAMP21-ALL and *IGH::DUX4* fusion, with detailed copy number profile by WGS showing the associated deletion around *ERG* (below).



Supplemental Figure 4. Copy number profile of seven variant cases.

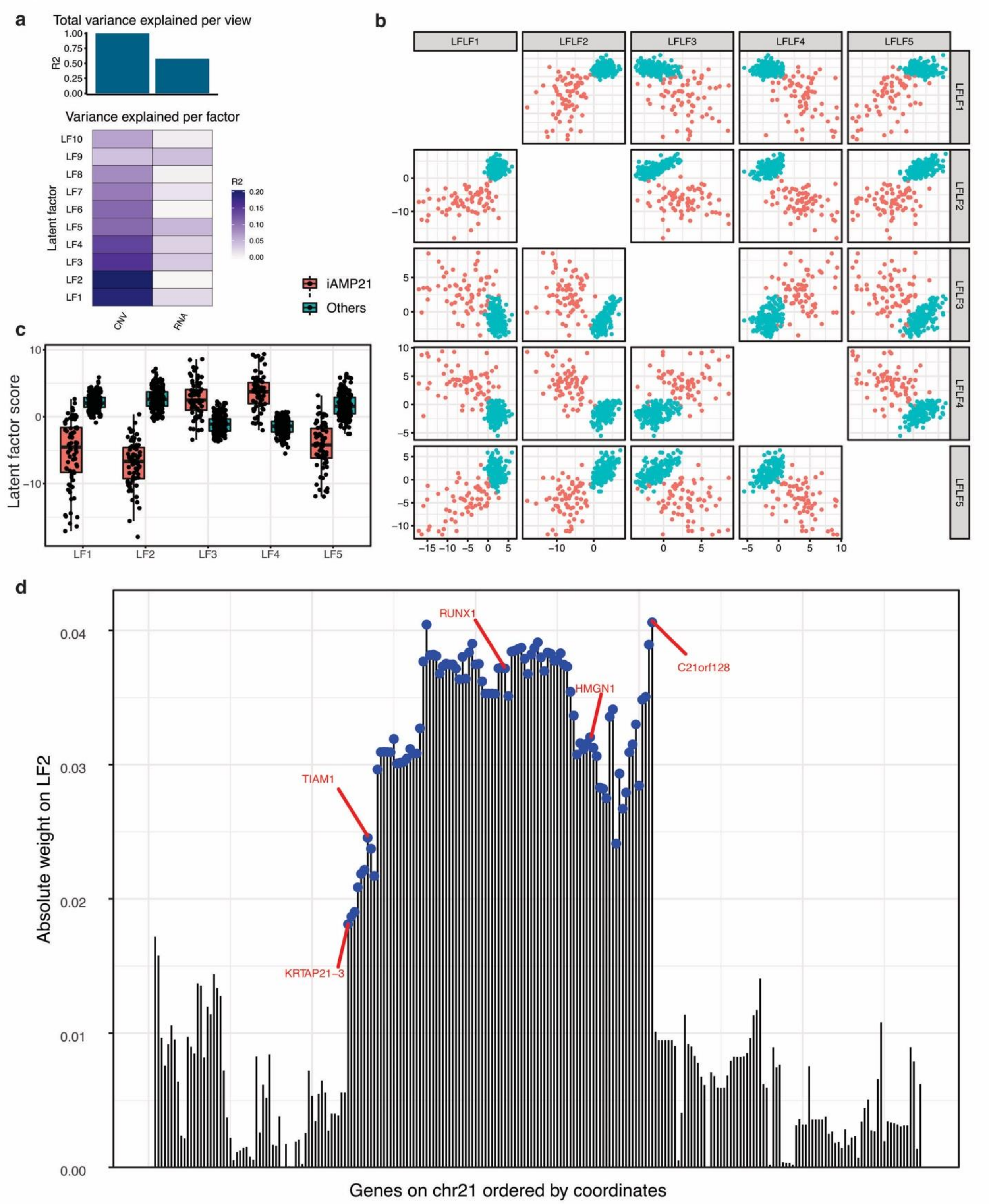
Copy number and rearrangement profiles of chromosome 21 are shown for seven variant iAMP21-ALL patients. They are defined as variants as they show no evidence of BFB and/or stepwise gain towards the highest region of copy number gain. Copy number is shown in the y-axis and the chromosome 21 ideogram is shown along the bottom of the plot, indicating breakpoint location. Rearrangements are separated based on their orientation.



Supplemental Figure 5. Expression pattern of genes on chromosome 21 across B-ALL subtypes.

(a) Heatmap showing the expression pattern of genes located around the common region of gain of chromosome 21 by rlog values, derived from whole transcriptome sequencing (WTS) data. Red and blue represent high and low levels of expression, respectively. Each row represents one patient and each column represents one gene. B-ALL subtype information is shown on the left. The ideogram for chromosome 21 and common region of overexpression (grey lines) are shown across the top of the heatmap. Genes are ordered based on their genomic coordinates on chromosome 21.

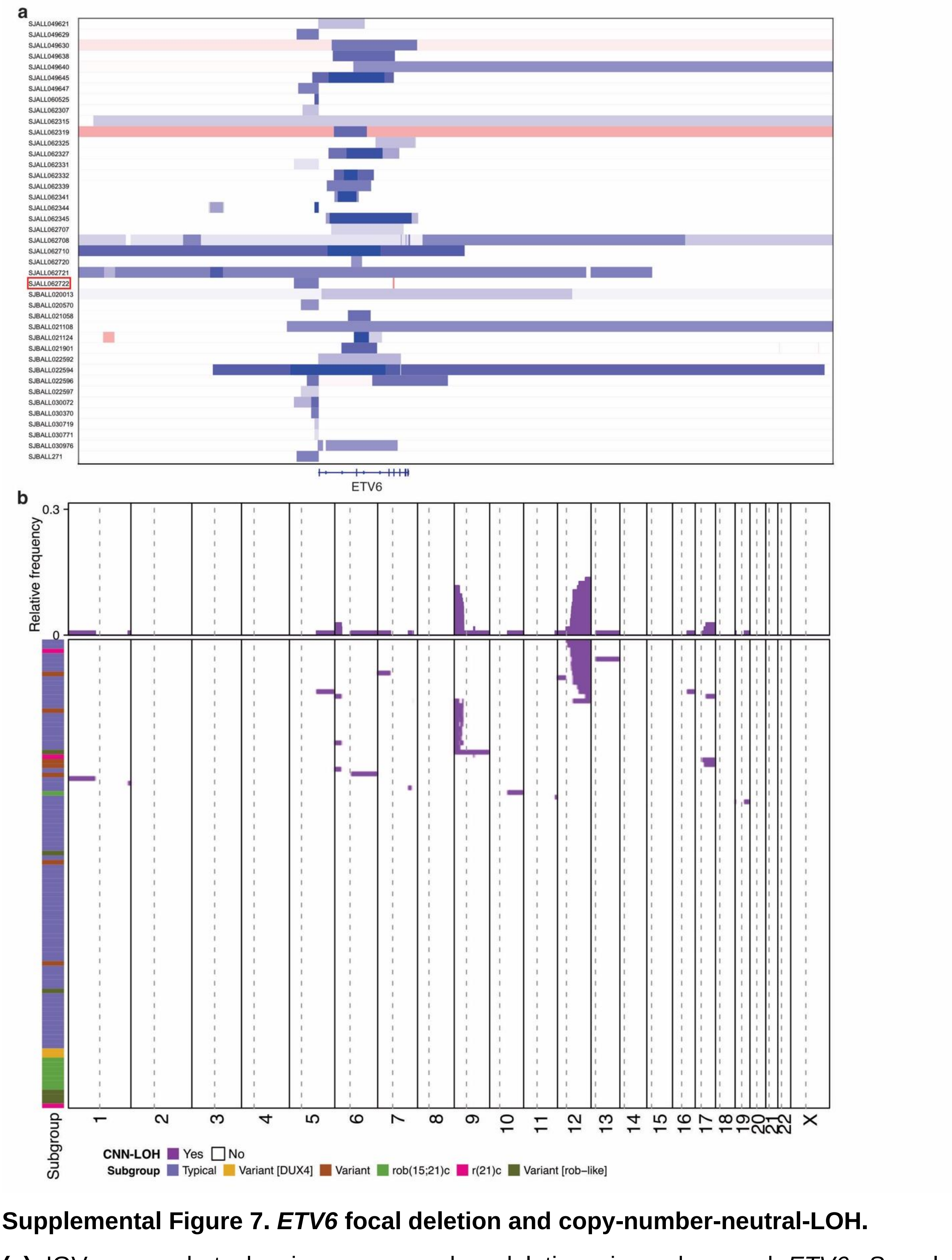
(b) Boxplot showing copy number profile of two regions: 28.3-29.3Mb and 35.8-36.8Mb (*RUNX1* region) on chromosome 21. Each dot represents a sample. Log_2 ratio of copy number was calculated between tumor and germline for each sample in each 0.1Mb window. The ideogram for chromosome 21, as well as GISTIC socres for copy number amplification are shown on the top.



Supplemental Figure 6. Unsupervised multi-omics factor integration analysis of genes on chromosome 21.

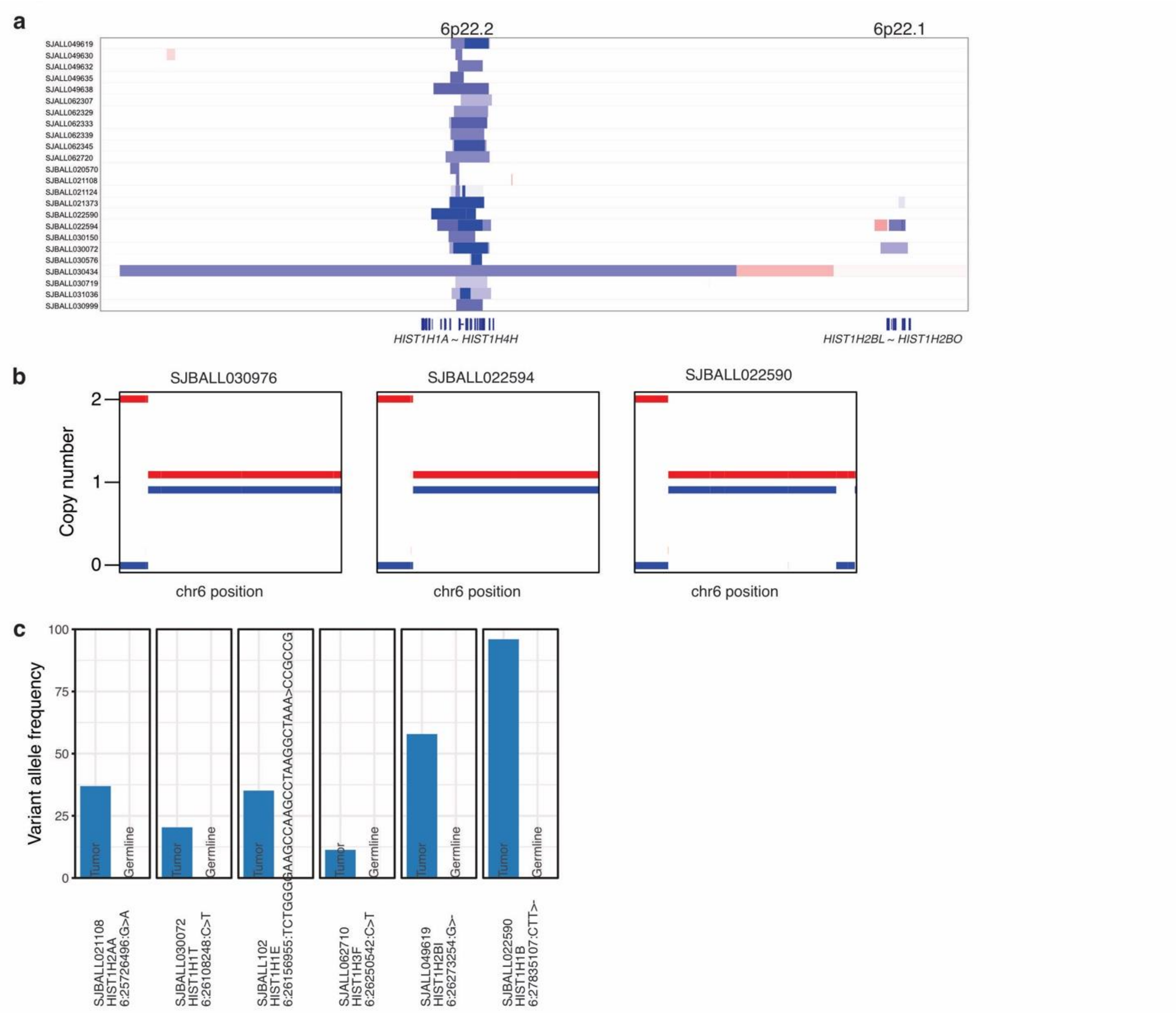
(a) Summary of variance in latent factors (LF1-LF10) derived from MOFA analysis across data types - CNV and expression (RNA). Variance explained by expression for LF2 is 0.

(b) Scatterplot showing cluster pattern of 70 iAMP21- and 179 non-iAMP21-high hyperdiploid-ALL patients in 2D space (pairwise combinations of the first five LFs, LFLF1-LFLF5). **(c)** Boxplot comparing latent factor scores between iAMP21- and non-iAMP21-ALL samples for the first five LFs, indicating the best LF2 separated iAMP21- and non-iAMP21-ALL samples (p-value = 2.0e-34). **(d)** Distribution of absolute loading weight on LF2 associated with CNV for each gene on chromosome 21 indicating an enrichment in the region from the *KRTAP21-3* to *C21orf128* genes.



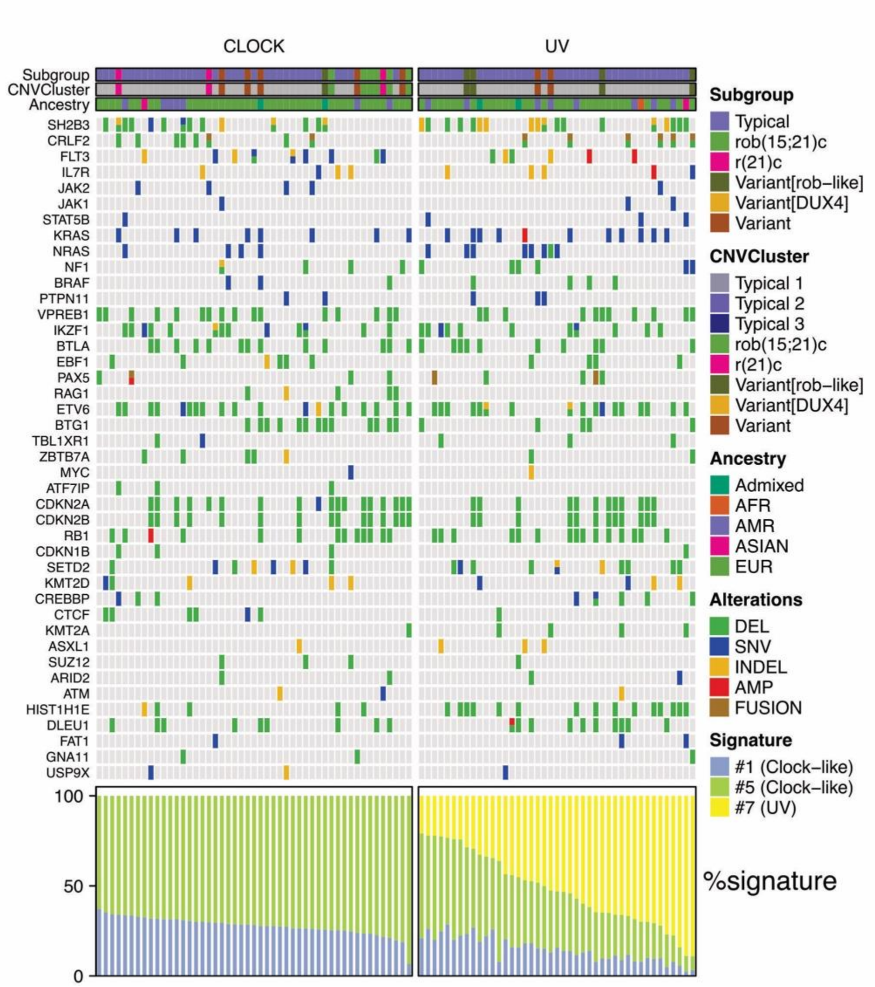
Supplemental Figure 7. *ETV6* focal deletion and copy-number-neutral-LOH.

(a) IGV screenshot showing copy number deletions in and around *ETV6*. Sample SJALL062722 (highlighted), clustered with *ETV6*::*RUNX1* ALL (see **Figure 1a**), showed an *ETV6* exon 1 deletion (part of a larger deletion, chr12:11,583,481-11,651,125) and exon 4 duplication (chr12:11,851,677-11,856,196). Likely these abnormalities are located on different alleles, representing biallelic abnormalities of *ETV6*. The patient has >40% APOBEC mutation signature, associated with *ETV6*::*RUNX1*/*ETV6*::*RUNX1*-like-ALL (**Supplemental Table 12**). **(b)** Heatmap showing genome wide copy-number-neutral-LOH (CN-LOH) for cases with high quality WGS (N=102). The presence of CN-LOH is shown in purple. The relative frequency for CN-LOH across all cases is shown at the top and iAMP21-ALL subgroup information is shown on the left.



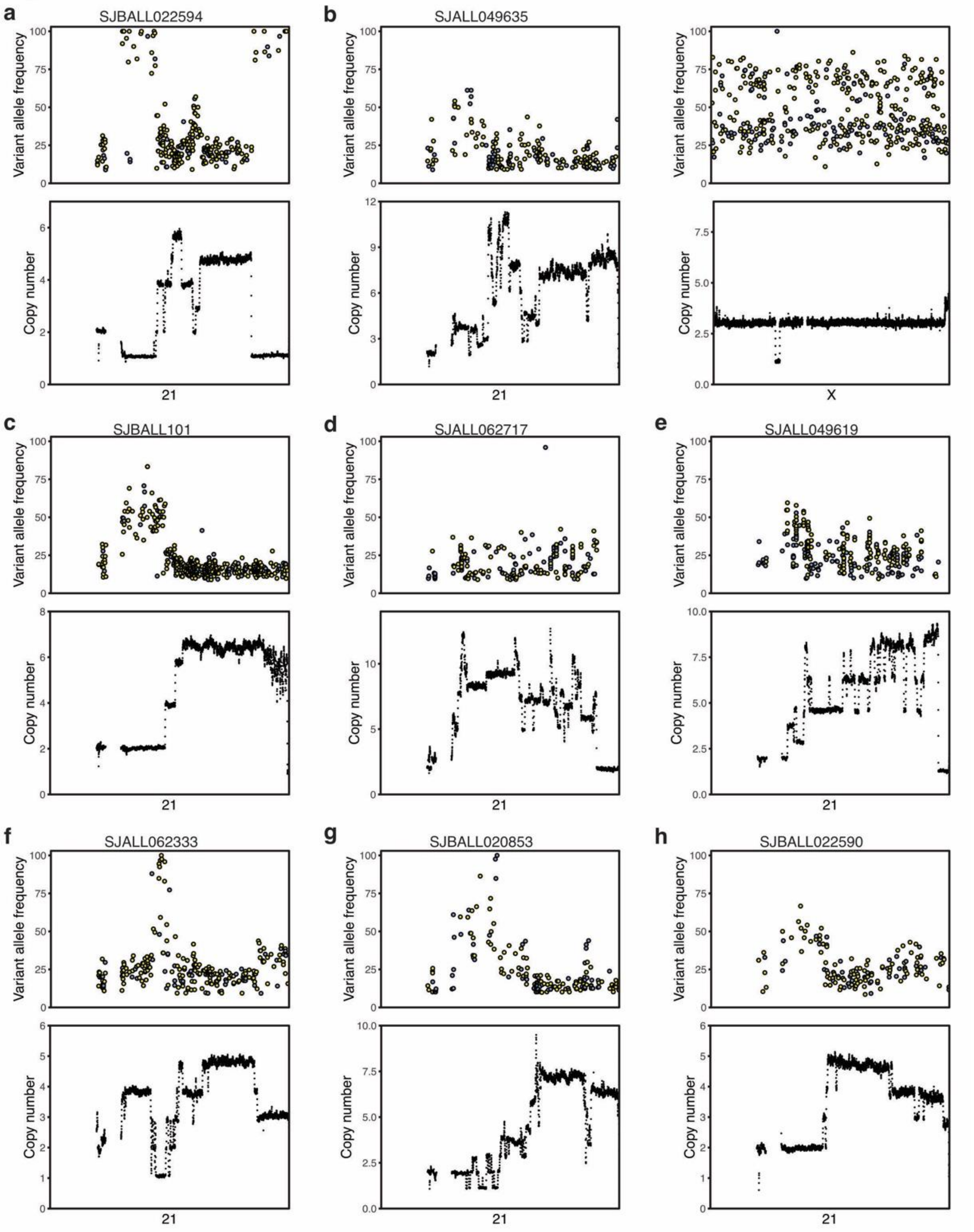
Supplemental Figure 8. Alterations of histone genes on 6p (related to Figure 3).

(a) IGV screenshot showing copy number deletions of two histone gene clusters on 6p. **(b)** Allele specific copy number of chromosome 6 in 3 cases with copy-number-neutral-LOH (CN-LOH). Red and blues lines represent two alleles. Lines are offset from discrete copy number values by ±0.1 for visual separation of the two alleles, showing copy number deletion of two histone gene clusters on 6p. **(c)** Barplots showing variant allele frequency for histone gene SNV and small indels in tumor and germline samples.



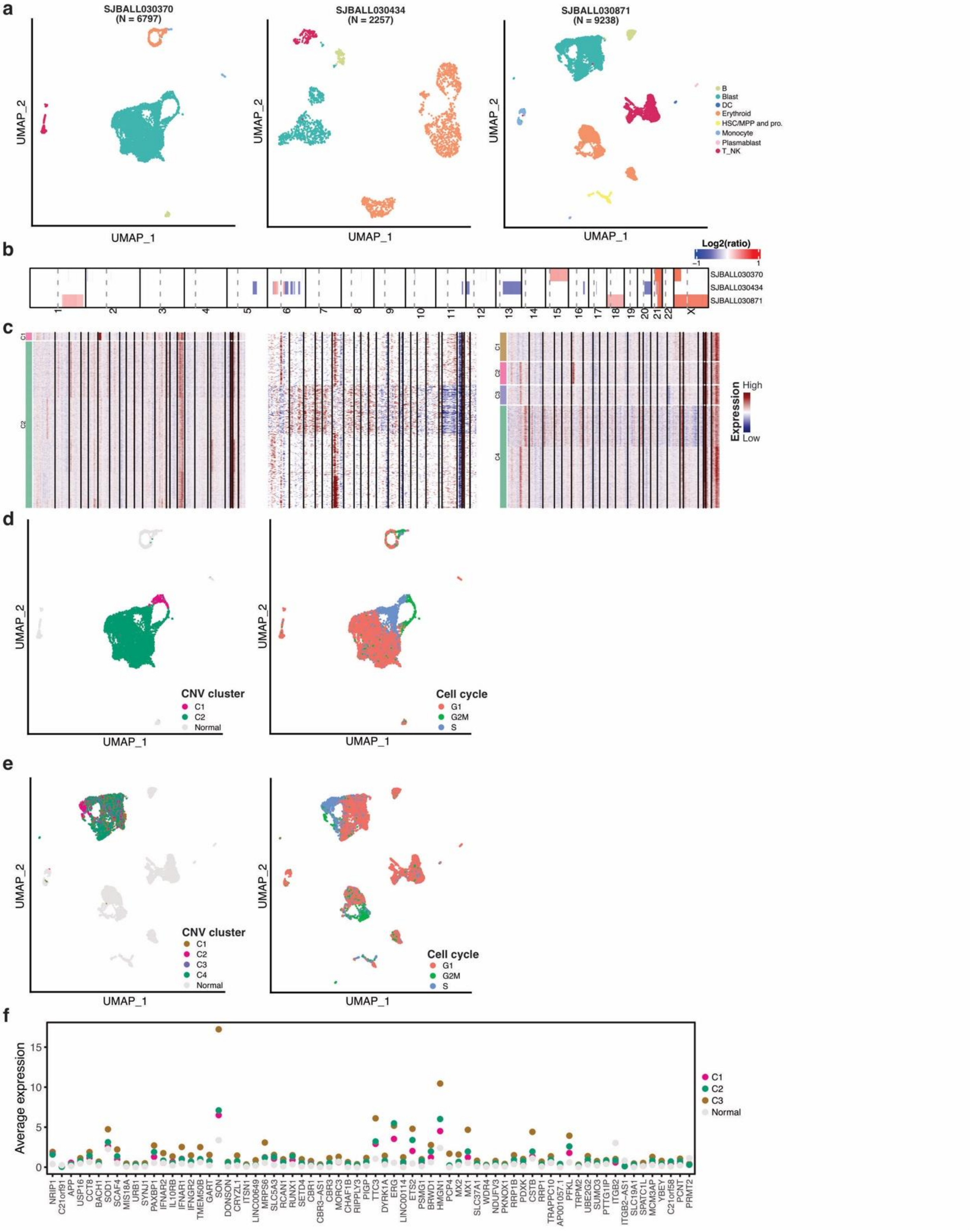
Supplemental Figure 9. Clinical and genetic features for cases with or without UV signature.

Oncoprint of recurrent and focal genomic changes, clinial features and subgroups for cases with or without UV signature. No significant differences were observed (**supplemental Table 13**).



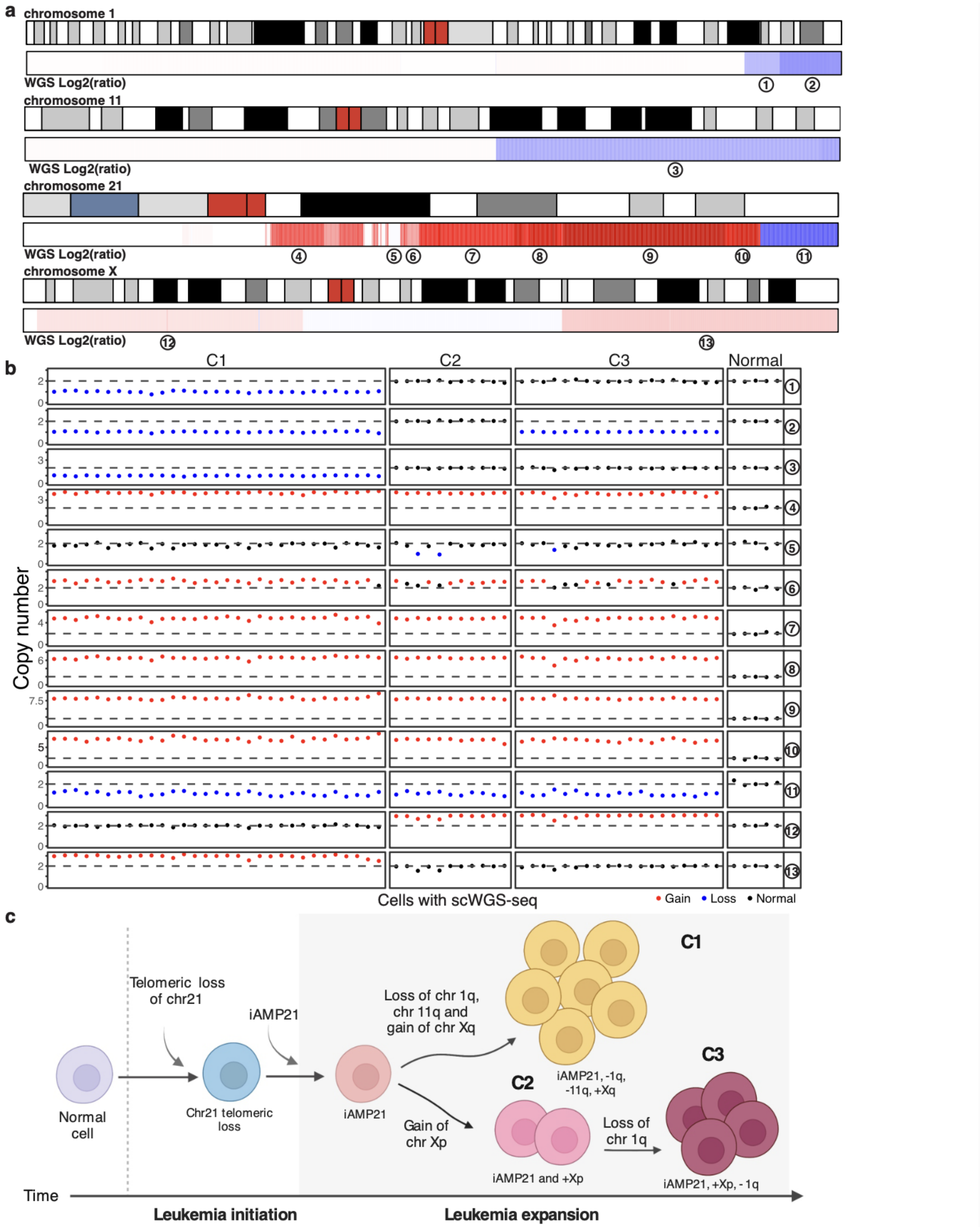
Supplemental Figure 10. Variant allele frequency and copy number profile for eight patients with >100 mutations on chromosome 21.

Scatterplot showing VAF distribution of somatic mutations on chromosome 21, colored by mutation signatures (upper panel; yellow for UV and blue for others), and the copy number profile of the iAMP21 chromosome (lower panel) in eight patients: **(a)** SJBALL022594, **(b)** SJALL049635, **(c)** SJBALL101, **(d)** SJALL062717, **(e)** SJALL049619, **(f)** SJALL062333, **(g)** SJBALL020853 and **(h)** SJBALL022590, respectively. Patient SJALL049635 has copy number gain of chromosome X with >100 UV mutations, which is also shown in **(b)**. The VAF of UV-induced mutations was consistently low in the regions of highest copy number of chromosome 21 and were therefore not represented on the amplified allele. This demonstrates that UV-induced mutations followed formation of the iAMP21 chromosome. Conversely, UV-induced mutations were observed on 1 of 3 alleles or 2 of 3 alleles of chromosome X in patient SJALL049635 **(b)**, demonstrating the UV-induced mutations occurred before gain of chromosome X.



Supplemental Figure 11. Clonal evolution of iAMP21-ALL using scRNA-seq.

(a) UMAP representation of the scRNA-seq data set in 3 iAMP21-ALL patients. Clusters of cells are colored by cell types. **(b)** Copy number profiles from WGS for the same 3 patients above. **(c)** scRNA-seq derived copy number profiles for blast cells. Two clusters were observed for SJBALL030370 with copy number gain on the p arm of chromosome 6 in C1. Four clusters were observed for SJBALL030871: C1 lacked copy number gain of the q arm of chromosome 1; C2 had copy number gain of the p arm of chromosome 6; C3 had copy number loss of the p arm of chromosome 17; other cells are shown in C4. Expression of genes on 6p varied between clusters, which was attributable to variation in expression of histone genes in this region during cell cycle progression. **(d)** UMAP representation of the scRNA-seq data for SJBALL030370 colored by the 3 copy number clusters (left) and cell cycle (right), respectively. **(e)** UMAP representation of the scRNA-seq data for SJBALL030871 colored by the 3 copy number clusters (left) and cell cycle (right), respectively. **(f)** Scatterplot showing average expression of the 70 differentially expressed genes along chromosome 21 in **Figure 5d** across cells from different copy number clusters and normal cells. 13 genes had significantly higher expression in C3 compared to C2, while none had lower expression, 42 genes had significantly higher expression in C1 compared to C2 and 30 genes had significantly higher expression in C1 compared to C3.



Supplemental Figure 12. Clonal evolution of SJBALL030072 using scWGS-seq.

(a) Heatmap showing copy number profiles of chromosomes 1, 11, 21 and X derived from WGS data of SJBALL030072. Ideograms of these chromosomes are shown on the top.

Thirteen segments with varying copy number patterns are labeled at the bottom. **(b)** Scatterplot showing copy numbers of the thirteen segments as illustrated in **(a)** for 66 cells with scWGS in SJBALL030072. These cells are assigned to different clones (C1, C2, C3 and normal) based on their copy number patterns. **(c)** Schematic representation of the clonal evolution model in sample SJBALL030072. Chromosome 21 telomeric loss and copy number gain of the iAMP21 chromosome occur first, followed by two evolutionary tracks: (1) loss of chromosome 1q, 11q and gain of chromosome Xq in C1. (2) gain of chromosome Xp in C2, and further loss of chromosome 1q in C3. Notably, region with loss of chromosome 1q in C3 (18.8Mb) is smaller than that in C1 (29.6Mb).

SUPPLEMENTAL REFERENCES

1. Andrews S. FastQC: A quality control tool for high throughput sequence data [online]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.
2. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
3. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2).
4. Mayrhofer M, Viklund B, Isaksson A. Rawcopy: Improved copy number analysis with Affymetrix arrays. *Sci Rep*. 2016;636158.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
6. Broad Institute. Picard Toolkit.: Broad Institute; 2019.
7. Broad Institute. GATK: Genome Analysis Toolkit; 2021.
8. Chen X, Gupta P, Wang J, et al. CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat Methods*. 2015;12(6):527-530.
9. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12(4):R41.
10. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333-i339.
11. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84.
12. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220-1222.
13. Jeffares DC, Jolly C, Hoti M, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 2017;814061.
14. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
15. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012;28(14):1811-1817.
16. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.
17. 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.
18. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214-218.
19. Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14(6):e8124.

20. Brady SW, Roberts KG, Gu Z, et al. The genomic landscape of pediatric acute lymphoblastic leukemia. *Nat Genet*. 2022;54(9):1376-1389.
21. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169.
22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
23. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-883.
24. Krijthe JJ. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.16. https://github.com/jkrijthe/Rtsne. 2015.
25. van der Maaten L. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*. 2014;153221-3245.
26. van der Maaten L, Hinton G. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*. 2008;92579-2605.
27. Gu Z, Churchman ML, Roberts KG, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet*. 2019;51(2):296-307.
28. Tian L, Li Y, Edmonson MN, et al. CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol*. 2020;21(1):126.
29. Nicorici D, Satalan M, Edgren H, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014 doi:10.1101/011650.
30. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*. 2019;20(1):213.
31. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.
32. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267-273.
33. Bansal V, Libiger O. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics*. 2015;164.
34. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
35. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*. 2005;21(20):3896-3904.
36. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101.
37. Gerstung M, Jolly C, Leshchiner I, et al. The evolutionary history of 2,658 cancers. *Nature*. 2020;578(7793):122-128.
38. MutationTimeR. https://github.com/gerstung-lab/MutationTimeR.
39. Li Y, Schwab C, Ryan SL, et al. Constitutional and somatic rearrangement of chromosome 21 in acute lymphoblastic leukaemia. *Nature*. 2014;508(7494):98-102.