# Genomic landscape of Down syndrome-associated acute lymphoblastic leukemia

## Supplemental Tables

### Supplemental Table 2. Genes/loci that are significantly altered identified by gistic2 and MutSigCV.

| gene/locus | cytoband | Significantly altered (CNA) | Significantly altered (SNV/indel) | gistic2 q-value (CNA) | MutSigCV p-value (SNV/indel) |
|---|---|---|---|---|---|
| *CDKN2A* | 9p21.3 | Yes | Yes | 1.31E-85 | 0.00148 |
| *VPREB1* | 22q11.22 | Yes | No | 2.29E-55 | >0.01 |
| *IKZF1* | 7p12.2 | Yes | Yes | 3.46E-30 | 9.8E-12 |
| *PAX5* | 9p13.2 | Yes | Yes | 5.5E-21 | 0 |
| *ETV6* | 12p13.2 | Yes | Yes | 1.8E-17 | 0.0000121 |
| 6p22.2 | 6p22.2 | Yes | No | 3.24E-33 | >0.01 |
| *SLX4IP* | 20p12.2 | Yes | No | 7.66E-09 | >0.01 |
| *EBF1* | 5q33.3 | Yes | No | 1.13E-08 | >0.01 |
| 6p22.1 | 6p22.1 | Yes | No | 1.48E-08 | >0.01 |
| *BTLA* | 3q13.2 | Yes | No | 0.000000201 | >0.01 |
| *EGLN1* | 1q42.2 | Yes | No | 7.8E-10 | >0.01 |
| *BTG1* | 12q21.33 | Yes | No | 0.00000172 | >0.01 |
| *ADD3* | 10q25.1 | Yes | No | 0.000036 | >0.01 |
| *RAG2* | 11p12 | Yes | No | 0.00024623 | >0.01 |
| *XBP1* | 22q12.1 | Yes | No | 0.00025127 | >0.01 |
| *CHD2* | 15q26.1 | Yes | Yes | 0.00026902 | 0.00402 |
| *KDM6A* | Xp11.3 | Yes | No | 0.00069954 | >0.01 |
| *ARMC2* | 6q21 | Yes | No | 0.03089 | >0.01 |
| *STAG2* | Xq25 | Yes | No | 0.037958 | >0.01 |
| *ZNF217* | 20q13.2 | Yes | No | 0.12276 | >0.01 |
| *TSC22D1* | 13q14.11 | Yes | No | 0.13507 | >0.01 |
| *SETD2* | 3p21.31 | Yes | No | 0.14153 | >0.01 |
| *CRLF2* | Xp22.33 | No | Yes | >0.2 | 0 |
| *KRAS* | 12p12.1 | No | Yes | >0.2 | 0 |
| *NRAS* | 1p13.2 | No | Yes | >0.2 | 0 |
| *JAK2* | 9p24.1 | No | Yes | >0.2 | 1.54E-14 |
| *FLT3* | 13q12.2 | No | Yes | >0.2 | 2.08E-08 |
| *IL7R* | 5p13.2 | No | Yes | >0.2 | 0.0000114 |
| *CREBBP* | 16p13.3 | No | Yes | >0.2 | 0.0000963 |
| *ZEB2* | 2q22.3 | No | Yes | >0.2 | 0.000477 |
| *USP9X* | Xp11.4 | No | Yes | >0.2 | 0.000862 |
| *GNB1* | 1p36.33 | No | Yes | >0.2 | 0.000902 |
| *PTPN11* | 12q24.13 | No | Yes | >0.2 | 0.00114 |
| *DOT1L* | 19p13.3 | No | Yes | >0.2 | 0.00141 |
| *SH2B3* | 19p13.3 | No | Yes | >0.2 | 0.002 |

**Supplemental Table 3. Association of somatic alterations and subtype.**

| | Number of patients with alteration | | | | | | | Fisher exact p-value | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| gene/loci | *CRLF2*-r (n=128) | *ETV6::RUNX1*-like (n=38) | C/EBPalt (n=26) | High hyperdiploid (n=10) | *IGH::IGF2BP1* (n=7) | *PAX5*alt (n=6) | Others (n=29) | *CRLF2*-r | *ETV6::RUNX1*-like | C/EBPalt | High hyperdiploid | *IGF2BP1::IGH* | *PAX5*alt | Others |
| *JAK2* | 64 | 0 | 0 | 0 | 0 | 0 | 0 | 4e-23 | 4e-06 | 3e-04 | 0.067 | 0.195 | 0.345 | 2E-04 |
| *CRLF2* | 16 | 0 | 0 | 0 | 0 | 1 | 1 | 0.001 | 0.085 | 0.231 | 1 | 1 | 0.372 | 0.704 |
| *SH2B3* | 3 | 0 | 0 | 1 | 0 | 0 | 4 | 0.483 | 0.614 | 1 | 0.288 | 1 | 1 | 0.008 |
| *IL7R* | 4 | 0 | 0 | 0 | 0 | 3 | 2 | 0.74 | 0.362 | 0.603 | 1 | 1 | 7e-04 | 0.291 |
| *VPREB1* | 38 | 16 | 2 | 3 | 1 | 2 | 5 | 0.473 | 0.046 | 0.018 | 1 | 0.677 | 0.667 | 0.267 |
| *PAX5* | 34 | 14 | 6 | 1 | 0 | 1 | 4 | 0.462 | 0.066 | 1 | 0.458 | 0.199 | 1 | 0.175 |
| *IKZF1* | 40 | 0 | 2 | 0 | 1 | 2 | 10 | 7e-04 | 3e-05 | 0.079 | 0.123 | 1 | 0.62 | 0.153 |
| *EBF1* | 22 | 3 | 0 | 0 | 0 | 0 | 2 | 0.002 | 0.778 | 0.09 | 0.608 | 1 | 1 | 0.751 |
| *BTG1* | 10 | 4 | 0 | 0 | 0 | 2 | 0 | 0.448 | 0.286 | 0.231 | 1 | 1 | 0.052 | 0.229 |
| *BTLA* | 11 | 2 | 0 | 0 | 1 | 1 | 2 | 0.325 | 1 | 0.229 | 1 | 0.401 | 0.355 | 1 |
| *RAG2* | 6 | 6 | 0 | 0 | 0 | 1 | 1 | 0.584 | 0.011 | 0.374 | 1 | 1 | 0.301 | 1 |
| *KRAS* | 15 | 11 | 6 | 1 | 1 | 2 | 1 | 0.152 | 0.024 | 0.248 | 1 | 1 | 0.226 | 0.093 |
| *NRAS* | 9 | 2 | 3 | 4 | 0 | 2 | 2 | 0.272 | 0.543 | 0.714 | 0.007 | 1 | 0.093 | 1 |
| *PTPN11* | 1 | 1 | 2 | 1 | 0 | 1 | 0 | 0.105 | 1 | 0.125 | 0.224 | 1 | 0.14 | 1 |
| *CDKN2A* | 42 | 8 | 10 | 0 | 2 | 6 | 3 | 0.205 | 0.331 | 0.263 | 0.037 | 1 | 5e-04 | 0.017 |
| *SLX4IP* | 15 | 4 | 1 | 1 | 0 | 0 | 0 | 0.108 | 0.752 | 0.709 | 0.601 | 1 | 1 | 0.148 |
| 6p22.2 | 30 | 11 | 5 | 0 | 4 | 0 | 4 | 0.645 | 0.29 | 0.808 | 0.123 | 0.045 | 0.343 | 0.342 |
| *SETD2* | 8 | 1 | 11 | 0 | 0 | 0 | 2 | 0.123 | 0.215 | 8e-07 | 0.606 | 1 | 1 | 1 |
| 6p22.1 | 13 | 1 | 1 | 0 | 1 | 0 | 2 | 0.091 | 0.322 | 0.702 | 1 | 0.419 | 1 | 1 |
| *CREBBP* | 10 | 0 | 3 | 2 | 1 | 0 | 2 | 0.812 | 0.085 | 0.419 | 0.163 | 0.419 | 1 | 1 |
| *KDM6A* | 6 | 0 | 8 | 0 | 1 | 0 | 4 | 0.092 | 0.051 | 2e-04 | 1 | 0.437 | 1 | 0.257 |
| *CHD2* | 10 | 2 | 0 | 1 | 0 | 0 | 4 | 0.624 | 1 | 0.229 | 0.521 | 1 | 1 | 0.127 |
| *DOT1L* | 9 | 1 | 2 | 1 | 0 | 0 | 0 | 0.262 | 0.698 | 0.635 | 0.428 | 1 | 1 | 0.375 |
| *ETV6* | 18 | 9 | 4 | 1 | 2 | 1 | 2 | 0.721 | 0.138 | 1 | 1 | 0.287 | 1 | 0.271 |
| *XBP1* | 11 | 0 | 1 | 0 | 0 | 0 | 3 | 0.114 | 0.137 | 1 | 1 | 1 | 1 | 0.399 |
| *TSC22D1* | 7 | 1 | 0 | 1 | 0 | 0 | 1 | 0.34 | 1 | 0.606 | 0.347 | 1 | 1 | 1 |
| *ZEB2* | 6 | 0 | 0 | 1 | 1 | 0 | 1 | 0.505 | 0.362 | 0.603 | 0.318 | 0.234 | 1 | 1 |
| *FLT3* | 5 | 0 | 11 | 1 | 0 | 0 | 3 | 0.017 | 0.05 | 2E-07 | 0.582 | 1 | 1 | 0.715 |
| *USP9X* | 14 | 0 | 0 | 0 | 0 | 0 | 3 | 0.011 | 0.083 | 0.229 | 1 | 1 | 1 | 0.435 |
| *ADD3* | 8 | 3 | 1 | 0 | 0 | 0 | 0 | 0.383 | 0.406 | 1 | 1 | 1 | 1 | 0.37 |
| *STAG2* | 6 | 3 | 2 | 0 | 2 | 0 | 0 | 0.777 | 0.433 | 0.635 | 1 | 0.047 | 1 | 0.375 |
| *ZNF217* | 6 | 0 | 0 | 0 | 0 | 1 | 1 | 0.286 | 0.614 | 1 | 1 | 1 | 0.183 | 1 |
| *ARMC2* | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 0.45 | 1 | 1 | 1 | 1 | 1 | 0.592 |
| *GNB1* | 3 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 0.007 | 1 |
| *EGLN1* | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0.349 | 1 | 0.365 | 0.155 | 1 | 0.095 | 1 |

**Supplemental Table 4. Summary of the cells analyzed in scRNA-Seq.**

| Condition | | WT control | WT *CEBPD* | Dp16 control | Dp16 *CEBPD* |
|---|---|---|---|---|---|
| Total #cells | | 10060 | 11246 | 17270 | 10031 |
| Analyzable cells post QC | | 8379 | 1225* | 16343 | 9907 |
| Clusters | | | | | |
| CLP | cluster 3 | 5078 | 162 | 30 | 12 |
| | cluster 10 | 325 | 24 | 40 | 42 |
| | cluster 12 | 39 | 0 | 31 | 0 |
| Pre-pro-B | cluster 5 | 701 | 28 | 812 | 136 |
| Pro-B | cluster 0 | 0 | 198 | 179 | 7294 |
| | cluster 4 | 0 | 112 | 76 | 2094 |
| | cluster 1 | 18 | 0 | 7220 | 166 |
| | cluster 2 | 20 | 0 | 6371 | 114 |
| | cluster 6 | 17 | 3 | 1360 | 39 |
| | cluster 11 | 0 | 0 | 167 | 3 |
| | cluster 9 | 831 | 12 | 6 | 2 |
| GMP | cluster 7 | 770 | 296 | 22 | 2 |
| | cluster 8 | 580 | 390 | 29 | 3 |
| Cell cycle (Pro-B only) | | | | | |
| G1 | | 63 | 205 | 5358 | 6985 |
| S | | 523 | 60 | 4341 | 859 |
| G2M | | 300 | 60 | 5680 | 1868 |

* for WT *CEBPD*, non-transduced cells were used to top up cells in library preparation. Only cells positive for either *CEBPD* or mCherry were kept for analysis.

**Supplemental Table 5. Comparison of alterations in *BCR*::*ABL1*-like and non-*BCR*::*ABL1*-like subtypes in CRLF2-r.**

| gene/loci | non-*BCR*::*ABL1*-like n=72 | *BCR*::*ABL1*-like n=26 | Fisher exact P-value |
|---|---|---|---|
| *JAK2* | 29 | 19 | 0.005726435 |
| *CDKN2A* | 26 | 8 | 0.810467727 |
| *IKZF1* | 12 | 20 | 5.69E-08 |
| *VPREB1* | 16 | 12 | 0.040775769 |
| *PAX5* | 20 | 7 | 1 |
| 6p22.2 | 23 | 1 | 0.003190229 |
| *EBF1* | 2 | 14 | 2.66E-08 |
| *CRLF2* | 11 | 4 | 1 |
| *SLX4IP* | 5 | 8 | 0.004728529 |
| *ETV6* | 9 | 3 | 1 |
| *KRAS* | 8 | 3 | 1 |
| *USP9X* | 2 | 9 | 7.48E-05 |
| *BTLA* | 4 | 6 | 0.020014329 |
| *CREBBP* | 7 | 2 | 1 |
| 6p22.1 | 8 | 0 | 0.105038446 |
| *BTG1* | 3 | 5 | 0.028980199 |
| *CHD2* | 2 | 5 | 0.013399053 |
| *NRAS* | 5 | 2 | 1 |
| *TSC22D1* | 2 | 5 | 0.013399053 |
| *XBP1* | 0 | 7 | 4.75E-05 |
| *ADD3* | 2 | 4 | 0.041020174 |

**Supplemental Table 6. Percentages of subtypes in DS-ALL and non-DS-ALL.** Subtypes significantly (P<0.0019, Bonferroni adjusted P<0.05) over- or under-represented in DS-ALL are colored in red or blue, respectively.

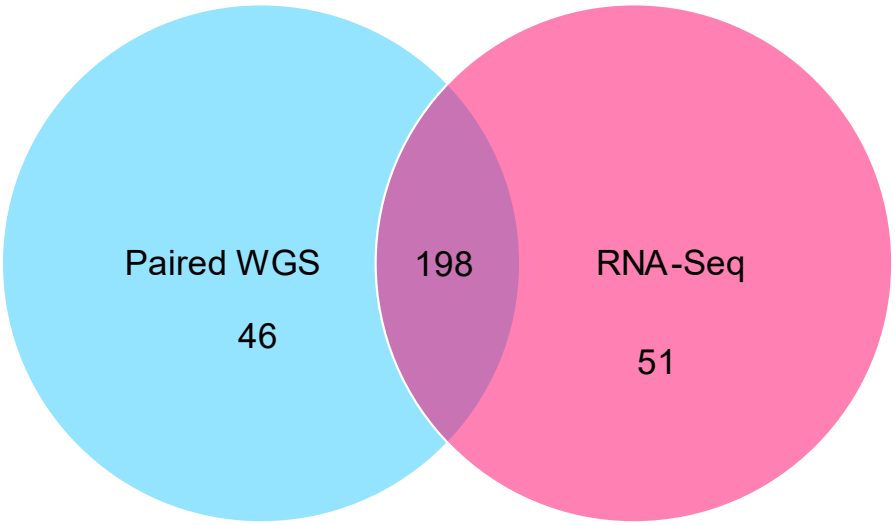
| Subtype | DS-ALL (N=295) | non-DS-ALL (N=2257)* | P |
|---|---|---|---|
| *CRLF2*-r | 54.24% | 6.03%** | $4.30\times10^{-88}$ |
| *ETV6*::*RUNX1* | 10.85% | 17.68% | 0.0028 |
| *ETV6*::*RUNX1*-like | 3.05% | 1.55% | 0.089 |
| C/EBP-altered | 10.51% | 0.13%*** | $9.1\times10^{-27}$ |
| High hyperdiploid | 4.41% | 23.00% | $5.5\times10^{-17}$ |
| *PAX5*alt | 2.71% | 5.80% | 0.028 |
| *IGH*::*IGF2BP1* | 2.71% | 0% | $2.9\times10^{-8}$ |
| *TCF3*::*PBX1* | 1.36% | 4.74% | 0.0056 |
| *BCR*::*ABL1*-like | 1.36% | 5.72% | $6.7\times10^{-7}$ |
| *BCR*::*ABL1* | 0.34% | 3.68% | $7.1\times10^{-4}$ |
| *KMT2A*-r | 0.34% | 3.72% | $4.6\times10^{-4}$ |
| *DUX4*-r | 0.34% | 4.25% | $1.3\times10^{-4}$ |
| *PAX5* P80R | 0.34% | 1.06% | 0.35 |
| *IKZF1* N159Y | 0.34% | 0.22% | 0.52 |
| iAMP21 | 0% | 5.67% | $2.1\times10^{-7}$ |
| Near haploid | 0% | 1.91% | 0.0074 |
| *MEF2D*-r | 0% | 1.55% | 0.028 |
| Low hypodiploid | 0% | 1.51% | 0.027 |
| *ZNF384*-r | 0% | 1.42% | 0.044 |
| *NUTM1*-r | 0% | 0.53% | 0.38 |
| *HLF*-r | 0% | 0.44% | 0.62 |
| *BCL2/MYC* | 0% | 0.13% | 1.00 |
| *KMT2A*-like | 0% | 0.09% | 1.00 |
| *ZNF384*-like | 0% | 0.09% | 1.00 |
| B-other | 7.12% | 9.04% | 0.33 |

*Subtypes classification of non-DS-ALL were obtained from Brady et al., Nature Genetics 54, 1376-89 (2022). **All the non-DS-ALL patients with *CRLF2* rearrangements were assigned as *CRLF2*-r. ***In non-DS-ALL, cases with rearrangements of C/EBP family genes are considered as C/EBPalt subtype (n=3).

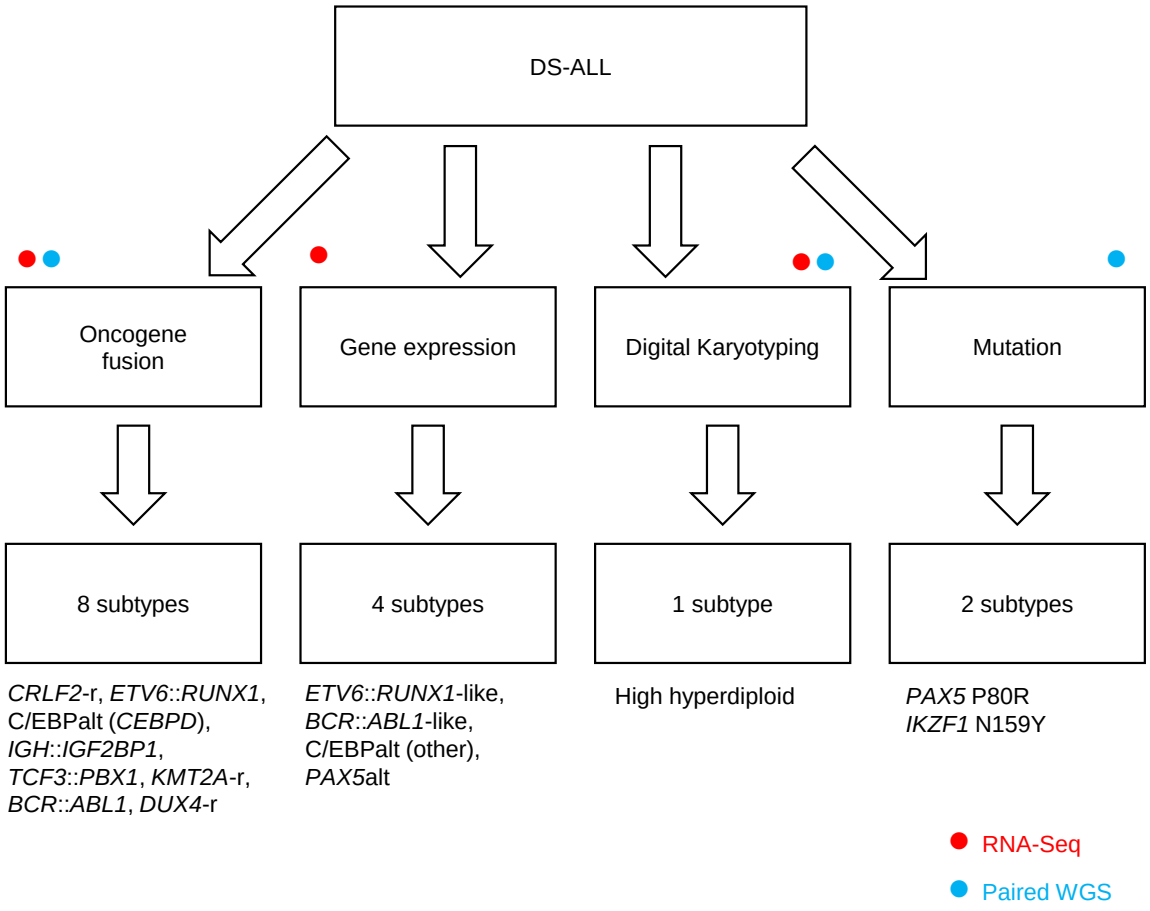
**Supplemental Table 7. Multivariate analysis of event-free and overall survival within *CRLF2*-r DS-ALL.**

| Variable | Event free survival | | Overall survival | |
|---|---|---|---|---|
| | Hazard ratio (95% CI) | P | Hazard ratio (95% CI) | P |
| *CRLF2*-r subgroup | | | | |
| non-*BCR*::*ABL1*-like | Reference | | Reference | |
| *BCR*::*ABL1*-like | **4.32 (1.71-10.92)** | **0.0020** | 2.38 (0.60-9.45) | 0.22 |
| NCI risk | | | | |
| Standard risk | Reference | | Reference | |
| High risk | 1.45 (0.61-3.43) | 0.40 | 4.52 (1.19-17.12) | 0.026 |
| EOI MRD | | | | |
| <0.01% | Reference | | Reference | |
| ≥0.01% | 1.73 (0.70-4.27) | 0.24 | 1.03 (0.27-4.00) | 0.96 |
| Rearrangement | | | | |
| *P2RY8* | Reference | | Reference | |
| *IGH* | 0.57 (0.18-1.79) | 0.34 | **0.07 (0.01-0.61)** | **0.016** |
| *IKZF1*del status | | | | |
| non *IKZF1*del | Reference | | Reference | |
| *IKZF1*del | 1.00 (0.39-2.60) | 1.00 | **5.70 (1.64-19.83)** | **0.0062** |

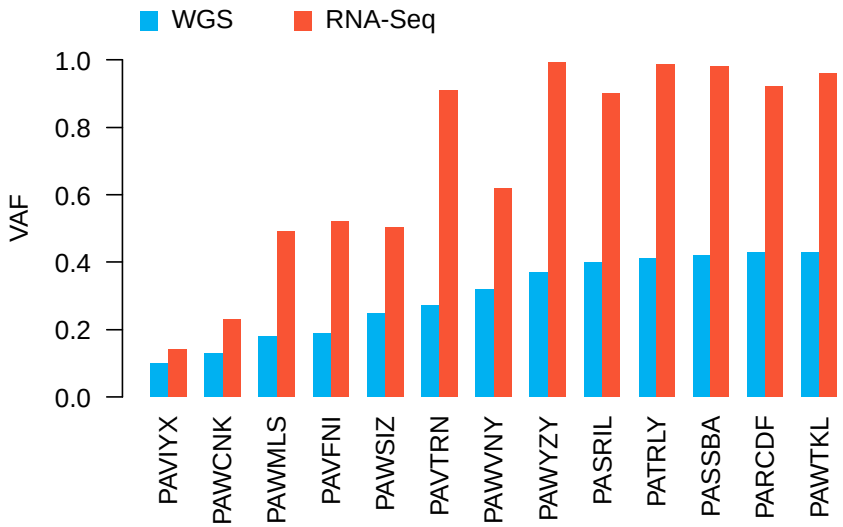
## Supplemental Figures



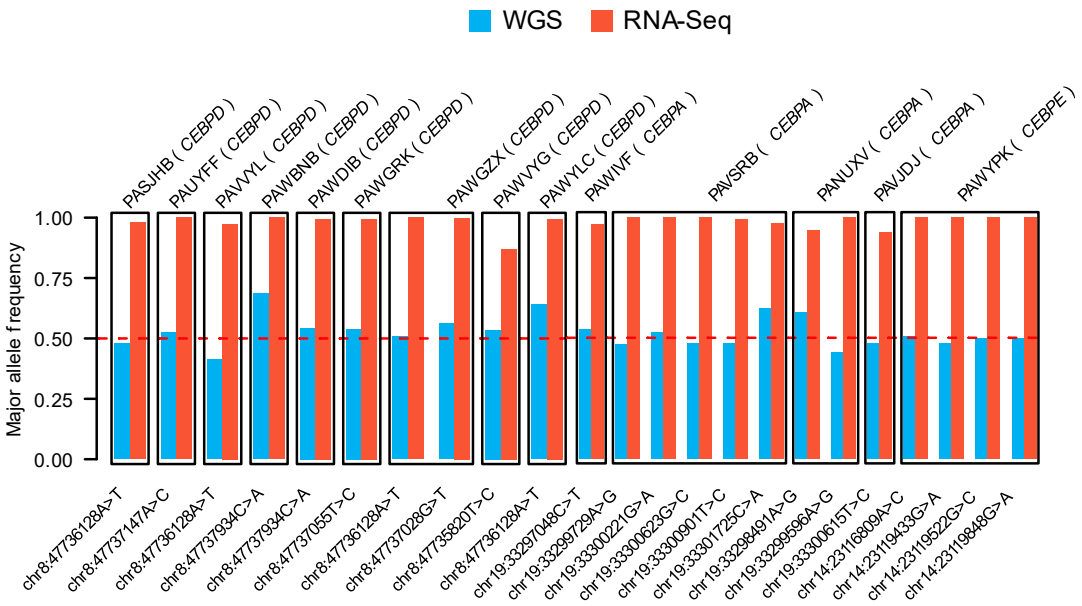

**Supplemental Figure 1. Next generation sequencing study of DS-ALL**. Venn diagram showing the number of patient samples subjected to whole genome sequencing (WGS) of paired leukemia and germline samples and RNA sequencing (RNA-Seq) of leukemia samples.



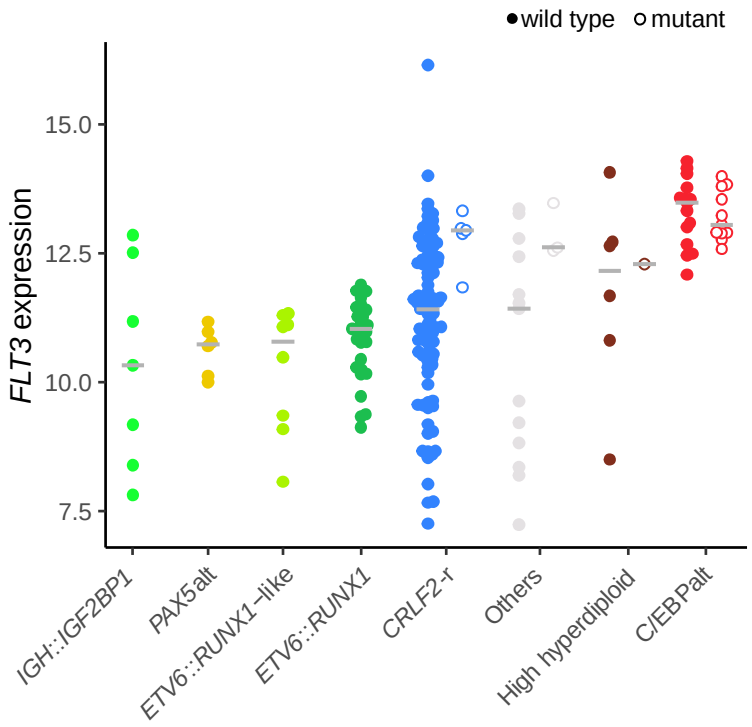

**Supplemental Figure 2. DS-ALL subtype classification workflow.**



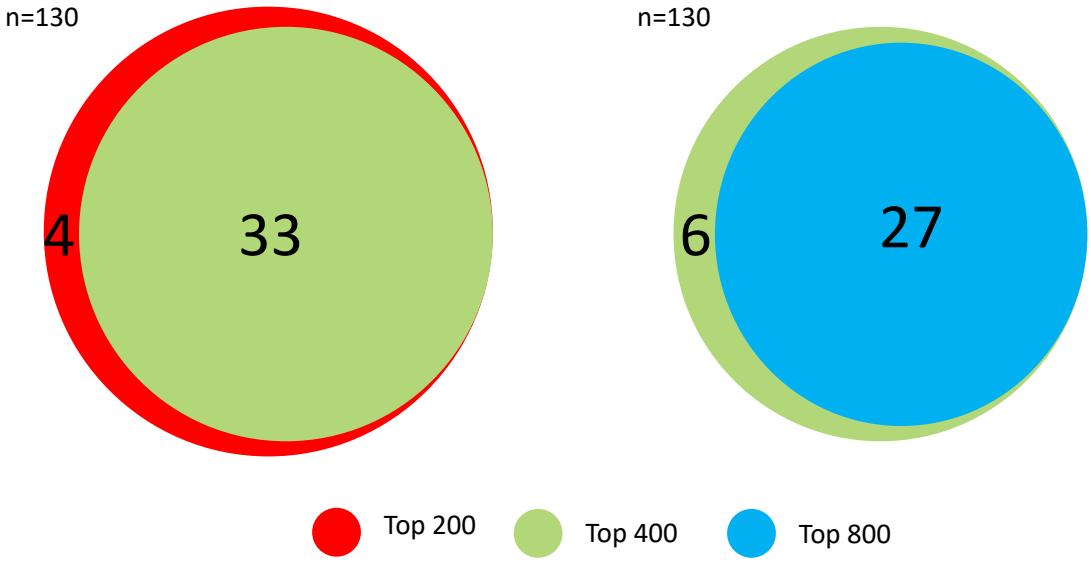

**Supplemental Figure 3. Allelic expression of *CRLF2* p.F232C mutation.** Comparison of the variant allele frequency of reads by tumor whole genome sequencing and RNA-seq.



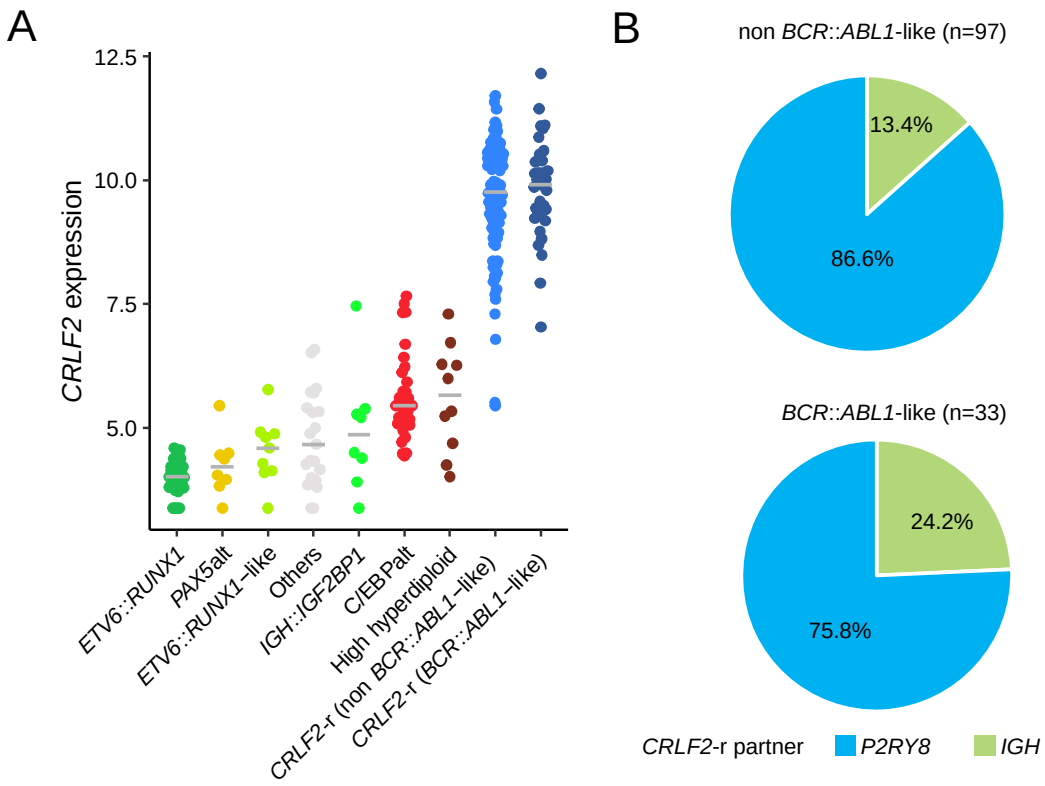

**Supplemental Figure 4. Major allele frequency (MAF), by WGS and RNA-Seq of tumor samples, of SNPs within 3000 bp of C/EBPalt cases with *CEBPD*, *CEBPA* or *CEBPE* alterations.** Only SNPs that were heterogeneous in WGS data, and had more than 10 reads coverage in both WGS and RNA-Seq are included. The major allele is defined as the allele with higher coverage (>0.5) in RNA-Seq. The red dashed line indicates allele frequency of 0.5. These SNPs had MAF near 0.5 in WGS but MAF near 0 or 1 by RNA-Seq, suggesting only one allele was expressed.



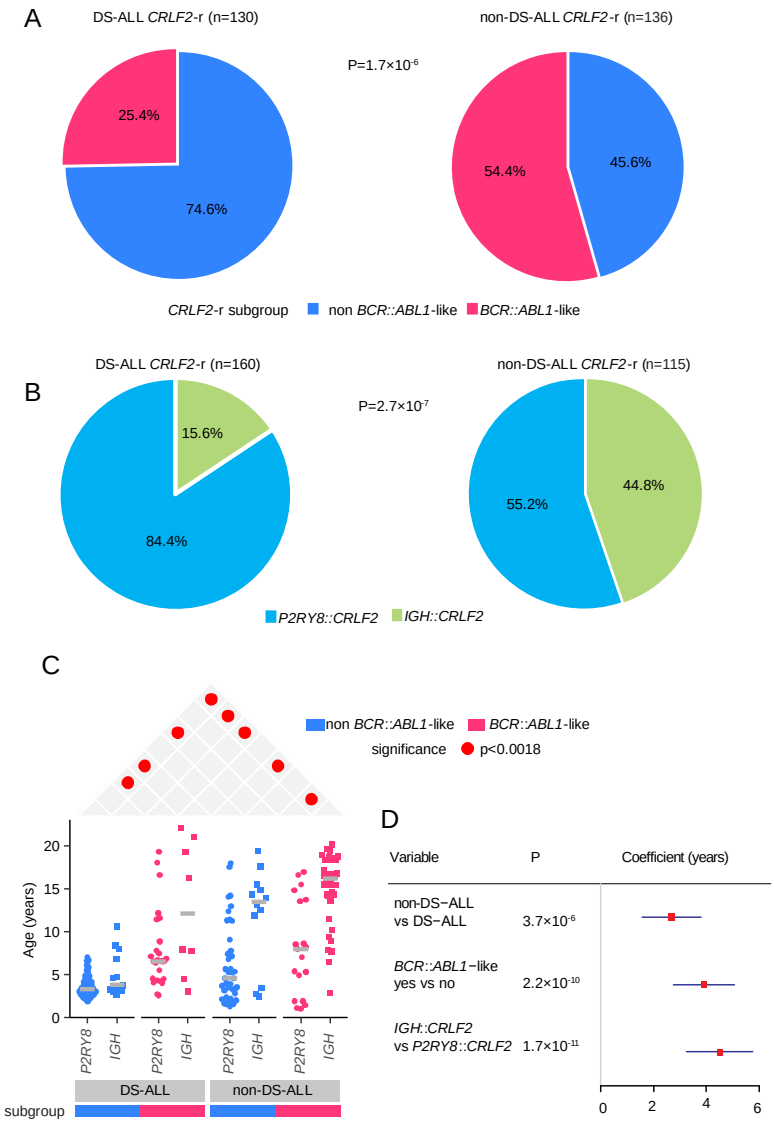

**Supplemental Figure 5. *FLT3* expression in DS-ALL subtypes with or without *FLT3* alterations.**



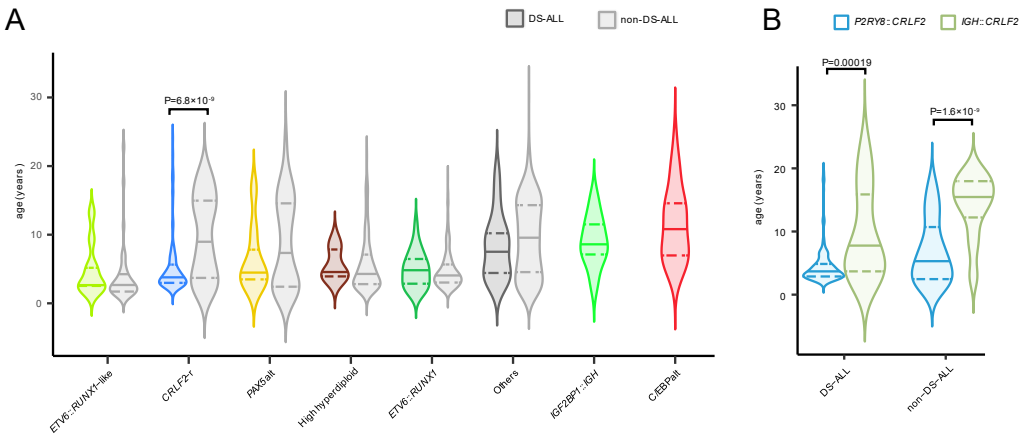

**Supplemental Figure 6. *CRLF2*-r *BCR*::*ABL1*-like cases identified with varied number of genes.** Number of *CRLF2*-r *BCR*::*ABL1*-like cases identified using unsupervised clustering of *CRLF2*-rearranged DS-ALL with varying number of top genes with the largest median absolute deviation.



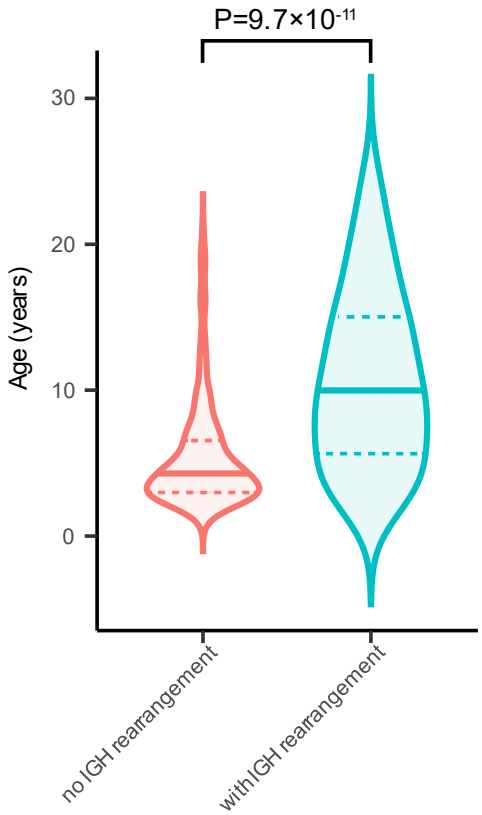

**Supplemental Figure 7. *CRLF2* expression and rearrangement in *BCR*::*ABL1*-like and non *BCR*::*ABL1*-like subgroups. A.** expression of *CRLF2* in subtypes showing that *BCR*::*ABL1*-like and non *BCR*::*ABL1*-like *CRLF2*-rearranged DS-ALL have similarly high expression of *CRLF2* (P=0.54). **B.** Proportions of *IGH*::*CRLF2* and *P2RY8*::*CRLF2* in the two sub-entities of *CRLF2*-rearranged DS-ALL.



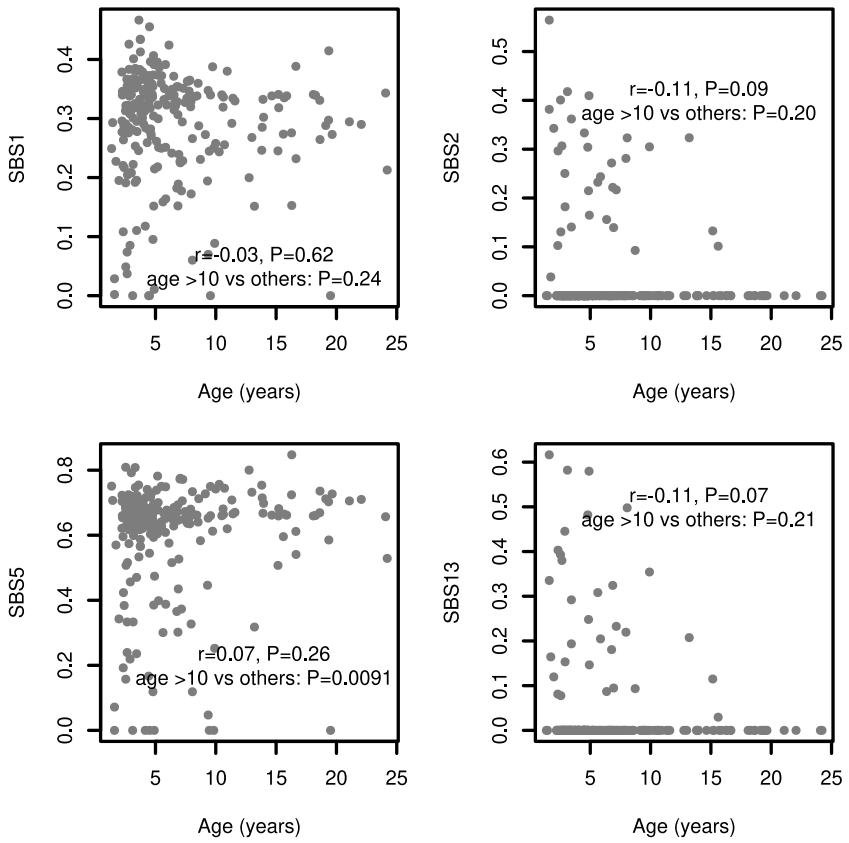

**Supplemental Figure 8. Comparison of the *CRLF2*-r subtype in DS-ALL and non-DS-ALL.** **A.** Frequencies of non *BCR*::*ABL1*-like and *BCR*::*ABL1*-like subgroups in *CRLF2*-r DS-ALL and non-DS-ALL. **B.** Frequencies of *CRLF2*-r partners in DS-ALL and non-DS-ALL. **C.** Distribution of age and statistical significance of difference in age of *CRLF2*-r patients according to rearrangement partner and *BCR*::*ABL1*-like status in DS-ALL and non-DS-ALL. Statistical significance is defined by Bonferroni adjusted P<0.05 (nominal P<0.0018). **D.** Association of DS status, *BCR*::*ABL1*-like gene expression signature, and *CRLF2* rearrangement partner with age at diagnosis in a multivariate linear regression model. Non-DS-ALL cases were compiled from Brady et al., Nature Genetics 54, 1376-89 (2022).



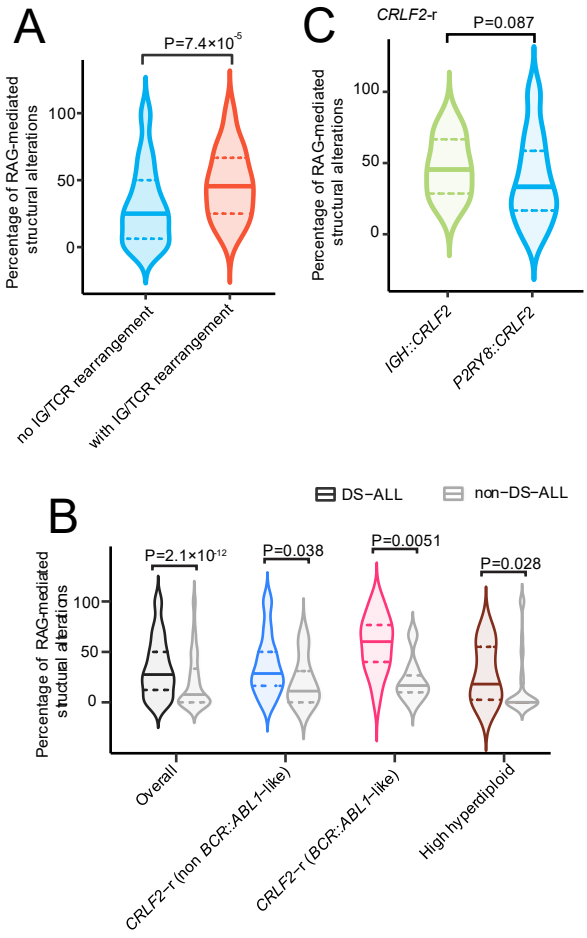

**Supplemental Figure 9. Comparison of age at diagnosis in DS-ALL and non-DS-ALL. A**. Comparison of age at diagnosis according to genomic subtypes in DS-ALL and non-DS-ALL. **B**. Comparison of age at diagnosis of *CRLF2*-r patients with *P2RY8*::*CRLF2* or *IGH*::*CRLF2* rearrangements in the DS-ALL and non-DS-ALL cohorts. The solid horizontal bars indicate the median of each group, and the dashed lines indicate the 25$^{th}$ and 75$^{th}$ percentiles.



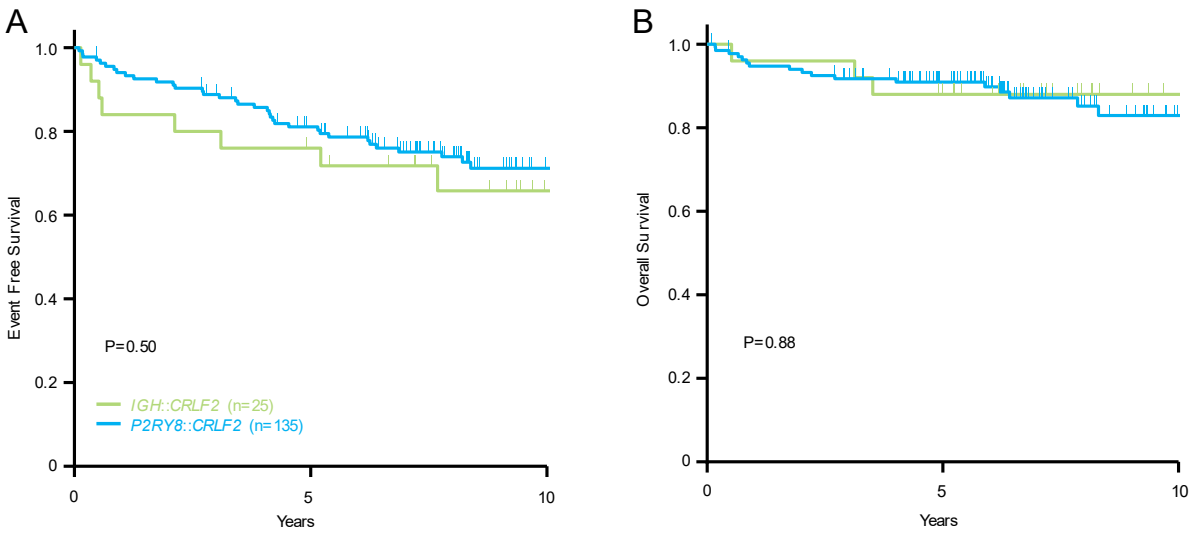

**Supplemental Figure 10. In DS-ALL, *IGH* rearrangement is associated with older age at diagnosis.**



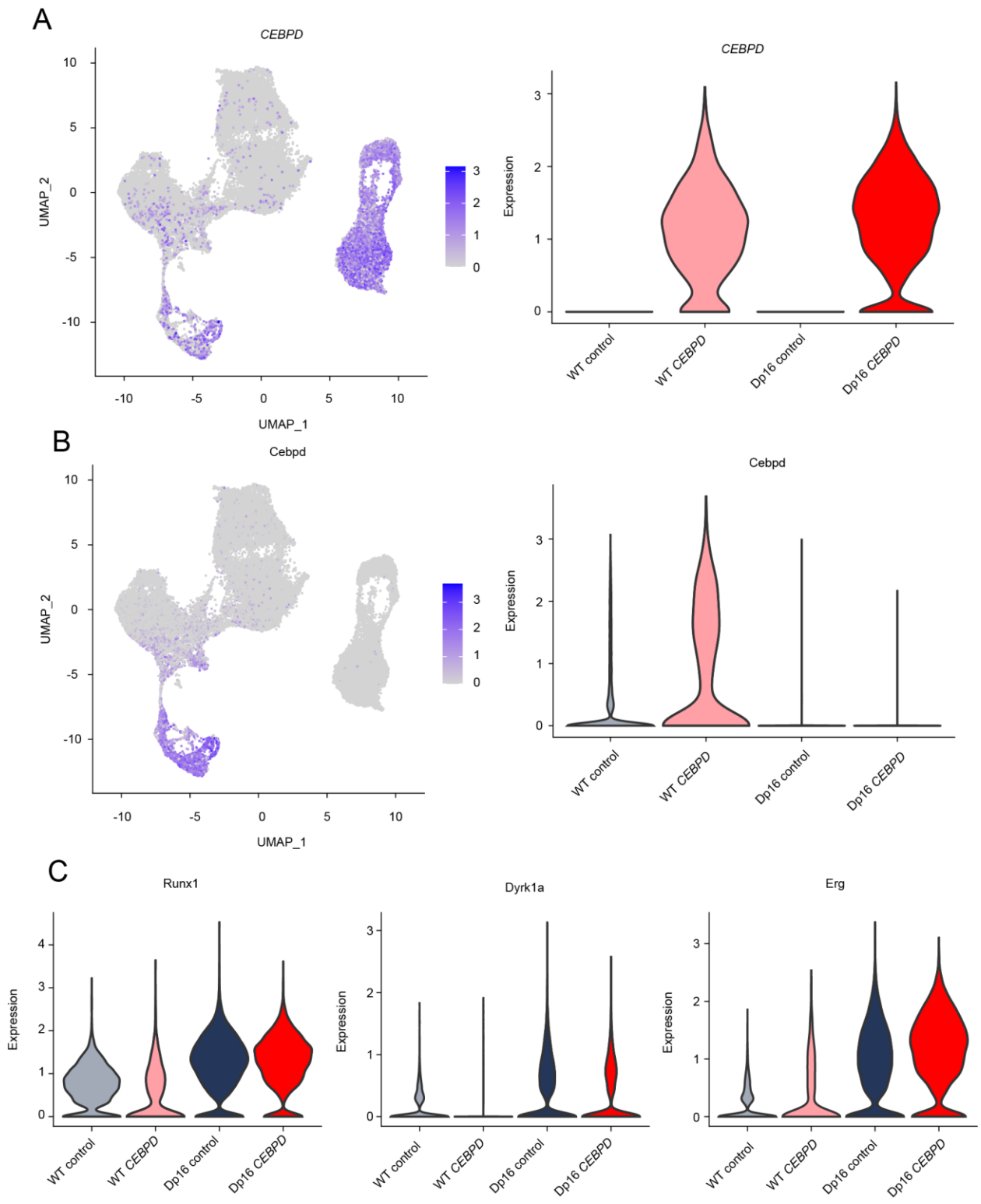

**Supplemental Figure 11. Association of age and mutation signatures.** Patient age is correlated with clock-like signature SBS5.




**Supplemental Figure 12. RAG-mediated structural alterations in DS-ALL. A.** Presence of *IGH* fusion is associated with the percentage of RAG-mediated structural alterations in DS-ALL. **B.** Percentage of RAG-mediated structural alterations in DS-ALL and non-DS-ALL subtypes. Comparisons were performed for *CRLF2*-r (*BCR*::*ABL1*-like and non *BCR*::*ABL1*-like), *ETV6*::*RUNX1*, *ETV6*::*RUNX1*-like, high hyperdiploid and *PAX5*alt subtypes and those with P<0.05 are shown. **C.** *IGH*::*CRLF2* and *P2RY8*::*CRLF2* had similar percentage of RAG-mediated structural alterations. Medians of each group are shown by solid horizontal bars and the 25$^{th}$ and 75$^{th}$ percentiles are indicated by dashed lines.




**Supplemental Figure 13. Outcomes of cases with *IGH*::*CRLF2* and *P2RY8*::*CRLF2* rearrangements.** *CRLF2* rearrangement partner (*IGH* or *P2RY8*) is not associated with differential **A**. event free survival or **B**. overall survival.




**Supplemental Figure 14. Gene expression of selected genes by scRNA-Seq.** scRNA-Seq data demonstrating expression of **A.** transduced human *CEBPD*, **B.** endogenous mouse *Cebpd*, and **C.** triplicated Hsa21 orthologues *Runx1*, *Dyrk1a*, and *Erg*.

## Supplementary methods

### WGS and RNA-Seq alignment and quality control

Alignment and quality control of pair-end WGS reads were performed using the Parabricks toolbox. This GPU-accelerated toolbox integrated bwa (v0.7.15) for alignment, GATK (v4.1.0.0) and samtools (v1.10) for processing and quality control. We required ≥75% of the reads to be mapped to the GRCh38 genome reference with duplication rate ≤20%, and cover ≥80% of coding region with 20× coverage. RNA-Seq reads were mapped to the GRCh38 genome reference using STAR[1]. In general, a minimum of 90 million raw reads and 60 million mapped reads with <15% mapped to ribosome genes were required.

### WGS SNV/indel, CNA and structural alteration calling

We applied an ensemble approach to call somatic mutations (SNV/indels) with multiple published tools, including Mutect2 (v4.1.2.0)[2], SomaticSniper (v1.0.5.0)[3], VarScan2 (v2.4.3)[4], MuSE (v1.0rc)[5], and Strelka2 (v2.9.10)[6]. Consensus calls by at least two callers were considered as confident mutations. Variants called by a single caller were rescued subsequently after variant quality review. The consensus call sets were next manually reviewed for the read depth, mapping quality, and strand bias to remove additional artifacts. The variant annotation was performed using Annovar[7].

Somatic copy number alterations (CNA) were determined by CONSERTING[8]. For somatic structural alteration calling, five callers were implemented, including Delly (v0.8.2)[9], Manta (v1.5.0)[10], Gridss (v2.5.0)[11], Lumpy (v0.3.0)[12] and novoBreak (v1.1.3rc)[13]. The structural alteration calls passed the default quality filters of each caller were merged using SURVIVOR (v1.0.7)[14] and genotyped by SVtyper (v0.7.1)[15]. The call sets were manually reviewed for the supporting soft-clipped and discordant read counts at both ends of a putative structural alteration site. A minimum of 5 supporting reads at both ends were required.

We used GISTIC2.0 (for CNA)[16] and MutSigCV v1.41 (for SNV/indels)[17] to identify significantly altered genes in DS-ALL, with q-value <0.2 (for GISTIC2.0) and p-value<0.01 (for MutSigCV), respectively.

### RNA-Seq fusion calling, gene expression analysis

Fusions were detected by FusionCatcher[18], STAR-Fusion[19] and Arriba[20]. Candidate fusions were manually reviewed and only the reliable ones were kept for analysis. Read count of genes were extracted from the alignment bam files using RSEM[21] and normalized using variance stabilizing transformation from the DESeq2 package[22]. UMAP was performed using the top 100 genes with the highest median absolute deviation in DS-ALL, with correlation coefficient as distance metric, and 15 as the size of the local neighborhood. Similar results were obtained using the top 200, 400, or 1000 genes. Chromosome-level copy number alterations were called from RNA-Seq data using the method described previously[23]. Differential expression analysis was performed using DESeq2 package[22].

## Subtype classification

Samples with *CRLF2*, *ETV6*::*RUNX1*, *IGH*::*IGF2BP1*, *TCF3*::*PBX1*, *KMT2A*, *BCR*::*ABL1*, and *DUX4* rearrangements were assigned to their respective subtypes. Because DS-ALL patients have constitutional trisomy 21, the high hyperdiploid subtype was defined as modal chromosome number >51, instead of the usual definition of >50 chromosomes[24]. We identified one case with a *PAX5* P80R mutation and another case with an *IKZF1* N159Y mutation using WGS data, and they were classified to the *PAX5* P80R and *IKZF1* N159Y subtypes, respectively. Using the gene expression profile of known subtypes in non-DS-ALL data, we predicted, in DS-ALL, the subtypes of *ETV6*::*RUNX1*-like, *BCR*::*ABL1*-like (non *CRLF2*-r), and *PAX5*alt. C/EBPalt was classified in two steps. First, cases with *CEBPD* rearrangements were assigned to C/EBPalt subtype. One case, which had high expression of *CEBPD* and clustered with cases positive of *CEBPD* rearrangements but had no detectable *CEBPD* rearrangement by RNA-Seq and no paired WGS data, was also classified as C/EBPalt subtype as well. In the second step, we performed hierarchical clustering based on the top 400 genes that were differentially expressed between the *CEBPD* rearranged cases and other known subtypes in DS-ALL (excluding the *CEBPD* gene itself). Cases clustered in the same branch of *CEBPD* rearranged cases were also classified as C/EBPalt subtype. Most of them harbored alterations of other C/EBP genes (*CEBPA* or *CEBPE*; see Figure 2D-E). The hierarchical clustering was performed using Ward algorithm and correlation coefficient as the distance metric. Identical results were obtained when using the top 200, 600, or 1000 differentially expressed genes in hierarchical clustering.

## *BCR*::*ABL1*-like and non-*BCR*::*ABL1*-like classification in *CRLF2*-r DS-ALL

*BCR*::*ABL1*-like and non-*BCR*::*ABL1*-like *CRLF2*-r DS-ALL were classified using unsupervised hierarchical clustering. The top 400 genes with the highest median absolute deviation in *CRLF2*-rearranged DS-ALL samples were used in hierarchical clustering, with Ward algorithm and correlation coefficient as distance metric. We varied the number of genes used to the top 200 or top 800, and obtained highly similar results, with 96.9% (n=126/130) and 95.4% (n=124/130) of cases assigned to the same group, respectively (**supplemental Figure 6**). We also performed supervised classification of the *BCR*::*ABL1*-like signature in *CRLF2*-rearranged DS-ALL, using a PAM[25] model trained on the non-DS-ALL data. Classification results are consistent with unsupervised hierarchical clustering in 88.5% (n=115/130; **Figure 4A**) of the cases.

## Cell lines

Lenti-X 293T cells were obtained from Takara Bio (Shiga, JP), and verified to be mycoplasma-negative. Lenti-X 293T cells were maintained in DMEM (Gibco, Waltham, MA) supplemented with 10% fetal bovine serum (FBS) (Gibco) and 1% penicillin-streptomycin (Life Technologies, Carlsbad, CA). OP9 stromal cells were provided by Margaret Goodell. OP9 cells were grown in Alpha MEM with 20% FBS, 1% penicillin-streptomycin, and 250 ng/mL Amphotericin B (Lonza, Basel, CH).

## Lentivirus generation

The mCherry lentiviral vector was generated as previously described[26]. The human *CEBPD* coding region was purchased from Genscript (Piscataway, NJ). Adaptor primers (TTCTCTAGGCGCCGGATGAGCGCCGCGCTCTTCAGCCT and TGCATGGATCCCTAGGTTACCGGCAGTCTGCTGTCCCGG) were added by PCR in the

presence of 5% DMSO, and the assembly was cloned into the mCherry backbone vector after EcoRI digestion, using the NEBuilder® HiFi DNA Assembly Cloning Kit (New England Biolabs, Ipswich, MA). CEBPD-mCherry or mCherry control vectors were transfected alongside packaging vectors pCAGkGP1.1R, pCAG4-RTR2, and pCAG VSVG into lenti-X 293T cells using lipofectamine 3000 from Invitrogen (Waltham, MA). Resulting lentiviruses were concentrated 100-fold using Lenti-X concentrator from Takara, and stocks were stored at -80C.

**Mouse bone marrow transduction**

Dp16 or WT mice (8-15 weeks old) were treated with 150 mg/kg 5-fluorouracil intraperitoneally (Sigma-Aldrich, St. Louis, MO). After 5 days, mice were sacrificed and HPC-enriched bone marrow was isolated from femurs. HPCs were cultured in StemSpan (Stem Cell Technologies, Vancouver, CA) containing 100 ng/mL mSCF (Shenandoah, Warwick, PA), 10 ng/mL mIL-3 (Peprotech, Rocky Hill, NJ), 10 ng/mL mIL-6 (Peprotech), and 250 ng/mL Amphotericin B for 48 hr. Cells were transferred in the same medium to 24 well non-tissue culture retronectin-coated plates (Takara). Lentiviral stocks at 1:50 dilution and 8 µg/mL polybrene (Sigma-Aldrich) were added, and cells were incubated for 72 hours.

## OP9 co-culture system and cell sorting

Each of the four conditions were lentivirally transduced and co-cultured simultaneously. After 72 hours of transduction, mouse HPC cultures were collected by trypsinization, rinsed with PBS, and plated onto sub-confluent T75 flasks of OP9 cells at 2-6X$10^6$ cells per flask, and supplemented with 10 ng/mL recombinant mouse IL-7 (Shenandoah, Warminster, PA) and 10 ng/mL recombinant mouse Flt3 ligand (Flt3L, Peprotech). Every 3-4 days, non-adherent cells were replated onto fresh sub-confluent OP9 cells. After 14 days of OP9 co-culture, non-adherent cells were harvested and viably frozen in IMDM containing 40% FBS and 15% DMSO. For cell sorting, samples were thawed and stained with 7AAD (Becton Dickinson, Franklin Lakes, NJ). mCherry+ and 7AAD- cells from each of the four conditions were collected with a FACSAria II cell sorter (Becton Dickinson) for single-cell RNA-Seq.

## Single-cell RNA-Seq and analysis

We targeted at least 10,000 cells for each of the four samples for library preparation. For the WT CEBPD sample, only ~2,000 cells were obtained, and non-transduced WT cells were used to top up this sample to fulfill the minimum cell number requirement in library preparation. The Single cell 5' Gene Expression Libraries were prepared according to Chromium Single Cell Immune Profiling Solution 5'v2 (10x Genomics, Pleasanton, CA). In brief, single cells, reverse transcription (RT) reagents, Gel Beads containing barcoded oligonucleotides, and oil were loaded on a Chromium controller (10x Genomics) to generate single cell GEMS (Gel Beads-In-Emulsions) where full length cDNA was synthesized and barcoded for each single cell. Subsequently the GEMS were broken and cDNA from each single cell was pooled. Following cleanup using Dynabeads MyOne Silane Beads, cDNA was amplified by PCR. The amplified cDNA was fragmented to optimal size and the 5' Gene Expression (GEX) library was generated via End-repair, A-tailing, Adaptor ligation and PCR amplification. Samples were sequenced on a NovaSeq 6000 at an average of ~500M reads/sample. Cellranger (v7.0.0) was used to align the reads to a reference combining mm10 genome and the transduced *CEBPD* and mCherry sequences, followed by downstream analysis using Seurat (v4.1.0)[27]. For WT *CEBPD* sample, only cells with

either *CEBPD* or mCherry detected were kept for analysis. Cells with extremely high or low number of genes or molecules detected (more than 3 times the median absolute deviation above or below the median, respectively), or with >20% reads from mitochondrial genes were removed. *CEBPD* gene were exclusively detected in WT *CEBPD* and Dp16 *CEBPD* samples, confirming the specificity in transduction (**Supplemental Figure 14**). Runx1, Dyrk1a and Erg, triplicated in Dp16 cells and well known for their role in hematopoietic differentiation, were overexpressed in Dp16 cells (**Supplemental Figure 14**). Clustering of the cells was performed on the shared nearest neighbor graph constructed using the first 30 principal components, with a resolution of 0.4. A differentiation stage was assigned to each cluster after manual review of gene expression of markers defined previously in the Immunological Genome Project[28]. Individual cells were also assigned to cell cycle phases using cell cycle marker genes[29].

## References


1. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
2. 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.
3. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;28(3):311-317.
4. 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.
5. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*. 2016;17.
6. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*. 2018;15(8):591-+.
7. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
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. 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.
10. 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.
11. Cameron DL, Schroder J, Penington JS, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res*. 2017;27(12):2050-2060.
12. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84.
13. Chong Z, Ruan J, Gao M, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods*. 2017;14(1):65-67.
14. 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;8:14061.
15. Chiang C, Layer RM, Faust GG, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*. 2015;12(10):966-968.

16. 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.
17. 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.
18. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014:011650.
19. Haas BJ, Dobin A, Stransky N, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*. 2017:120295.
20. Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res*. 2021;31(3):448-460.
21. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
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. Barinka J, Hu Z, Wang L, et al. RNAseqCNV: analysis of large-scale copy number variations from RNA-seq data. *Leukemia*. 2022;36(6):1492-1498.
24. O'Connor D, Enshaei A, Bartram J, et al. Genotype-Specific Minimal Residual Disease Interpretation Improves Stratification in Pediatric Acute Lymphoblastic Leukemia. *J Clin Oncol*. 2018;36(1):34-43.
25. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*. 2002;99(10):6567-6572.
26. Li Y, Yang W, Devidas M, et al. Germline RUNX1 variation and predisposition to childhood acute lymphoblastic leukemia. *J Clin Invest*. 2021.
27. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587 e3529.
28. Painter MW, Davis S, Hardy RR, Mathis D, Benoist C, Immunological Genome Project C. Transcriptomes of the B and T lineages compared by multiplatform microarray profiling. *J Immunol*. 2011;186(5):3047-3057.
29. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189-196.