T cell acute lymphoblastic leukemia (T-ALL) is an aggressive hematological malignancy of T cell progenitors, known to be a heterogeneous disease in pediatric and adult patients. Here we attempted to better understand the disease at the molecular level based on the transcriptomic landscape of 707 T-ALL patients (510 pediatric, 190 adult patients, and 7 with unknown age; 599 from published cohorts and 108 newly investigated). Leveraging the information of gene expression enabled us to identify 10 subtypes (G1–G10), including the previously undescribed one characterized by GATA3 mutations, with GATA3–R26Q capable of affecting lymphocyte development in zebrafish. Through associating with T cell differentiation stages, we found that high expression of LYL1/LMO2/SPI1/HOXA (G1–G6) might represent the early T cell progenitor, pro/pre/cortical/cortical stage with a relatively high age of disease onset, and lymphoblasts with Tlx3/Tlx1 high expression (G7–G8) could be blocked at the cortical/post-cortical stage, while those with high expression of NKX2-1/TAL1/LMO1 (G9–G10) might correspond to cortical/postcortical/mature stages of T cell development. Notably, adult patients harbored more cooperative mutations among epigenetic regulators, and genes involved in JAK-STAT and RAS signaling pathways, with 44% of patients aged 40 y or above in G1 bearing DNMT3A/IDH2 mutations usually seen in acute myeloid leukemia, suggesting the nature of mixed phenotype acute leukemia.

T-ALL | RNA sequencing | molecular subtyping | mutation | T cell differentiation stages

T cell acute lymphoblastic leukemia (T-ALL) is characterized by malignant transformation and proliferation of T cell progenitors (1, 2), accounting for 10 to 15% of pediatric and 20 to 25% of adult ALL cases (2, 3). Therapeutic progress has led to a gradual improvement in clinical outcomes, with a curable rate achieving up to 90% in children but much lower rate (60%) in adults (4). Recent advances in high-throughput genomic technologies have spurred the cohort-scale genetic analysis for identifying recurrent genetic abnormalities in T-ALL (5, 6). Recurrent genetic events in lymphoid neoplasia have been reported to cooperatively induce malignant transformation of normal thymocytes through transcriptional deregulation, leaving traces in the form of specific expression patterns (7–9). In most T-ALL patients, genetic abnormalities often involve genes encoding transcription factors, as highlighted by lymphoblastic leukemia-associated factors (e.g., LYL1, LMO1/2, TLX1/3, NKX2-1, and TAL1/2) (10). These genes may exhibit aberrant expression levels, when structural variations involve the T cell receptor gene (TCR) or enhancer regions from other partner genes: for example, the transcription factor gene TAL1 in the STIL-TAL1 deletion (11). Most commonly seen in early T cell precursor ALL (ETP-ALL) patients are the SET-NUP214 fusion and fusions involving the gene NUP98 (12). In addition to gene fusions and rearrangements, mutations are observed in over 90% of T-ALLs (9). Driver mutations can occur in genes essential for regulating T lymphocyte development, to name a few: the NOTCH signaling (NOTCH1, FBXW7, NOTCH3), the JAK-STAT signaling (JAK3, STAT5B), the PI3K-AKT-mTOR (mammalian target of rapamycin) signaling (PIK3R1, PIK3CD), epigenetic regulators (PHF6, USP7), PRC2 complexes (EZH2, SUZ12), and transcription factors (BCL11B, ETV6, GATA3). Mutations can also occur in genes essential for modulating cell proliferation and differentiation, such as the RAS signaling (KRAS, NRAS), cell cycle, or apoptosis-related factors (CDKN2A, CDKN3), and translational regulators (RPL10, RPL5). In staging of T cell maturation, T-ALL can be arrested at different stages of T cell development. Based on the clinical immunophenotypes, T-ALL has been

Significance

We provide transcriptomic insights into differences between pediatric and adult T cell acute lymphoblastic leukemia (T-ALL) patients through an international collaborative effort integrating RNA-sequencing data of 707 patients. Ten subtypes were identified, each characterized by distinct gene mutation profiles and dysregulated expression signatures of leukemogenic factors, and associated with T cell development stages. Adult T-ALL tends to have characteristics of early T cell precursor ALL, mostly corresponding to the mixed phenotype acute leukemia, whereas pediatric T-ALL shows a wide spectrum of aberrant molecular features, from early T cell precursor to mature T cell compartments. Our findings have important implications for disease mechanism of T-ALL that differs between pediatric and adult patients, facilitating further refined targeted therapy.
subclassified into ETP-ALL, pro/precortical, cortical, postcortical, and mature T-ALL (5, 13). Notably, the recurrent genetic alterations show a degree of association with T cell development stages (14, 15), while combining both information from genetic alterations and differentiation arrests for improved subtyping has not yet been reported.

It is well recognized that T-ALL display a large variability of clinical and genetic features between pediatric and adult patients (9). Previously we described the landscape of B cell precursor-ALL molecular abnormalities, proposing 14 distinct subtypes based on RNA-sequencing (RNA-seq) data of 1,223 patients (7). By analogy, we here aimed to elucidate genetic/transcriptomic alterations alteration in T-ALL, in particular genomic insights into differences between pediatric and adult patients. We aggregated the evidence from RNA-seq data of 707 T-ALL patients (510 pediatric, 190 adult, and 7 with unknown age), including 599 obtained from six published international public cohorts (cohorts 1 to 6) (5, 9, 16–19) and 108 newly contributed from two more centers of excellence in China (cohorts 7 and 8) (SI Appendix, Table S1 and Dataset S1). All RNA-seq data were uniformly preprocessed and subjected to integrative analysis using a well-established pipeline (Fig. 1A) (7, 20, 21), generating resources on gene expression and genetic alterations.

Results

Discovery of Gene Fusions and Characterization of T-ALL Subtypes Based on RNA-seq Data from 707 Cases. We carried out an integrated analysis combining RNA-seq data from multiple
cohorts totaling 707 T-ALL patients (Fig. 1A). Based on the information of gene expression with batch effects corrected across cohorts (SI Appendix, Fig. S1 A and B), 10 distinct subtypes were identified using a graph-based semisupervised classification approach (SI Appendix, SI Materials and Methods). The classification of each subtype was evident in both visual inspection and robust analysis using random forest (SI Appendix, Fig. S1 C–G). The tightness of patient groupings was illustrated using the two-dimensional representation determined by t-distributed stochastic neighbor embedding (tSNE), complemented with subtype-specific expression illustrations using a suprahexagonal map, collectively revealing the (dis)similarity among subtypes (Fig. 1B).

We next characterized subtypes regarding demographic information, gene fusion events, and dysregulated expression of leukemogenic factors (Fig. 1 C and D and Dataset S2). In G1, G2, and G6, over 50% of patients were adults, representing the highest proportion of adult patients among all subtypes (Fig. 1C). G1 was associated with SET-NUP214 and NUP98 fusions and the elevated expression of LYL1 (participating in lymphogenesis) (22), LMO2 (perturbing T-cell differentiation) (23), SPI1 (expressing in ETPs and essential for the normal hematopoiesis) (24), MEF2C (activated in T-ALL) (25), and HOXA family genes (Fig. 1D, expression panel). Notably, ETP-ALL patients were mainly classified into G1 as compared to other subtypes (42 of 59 vs. 7 of 215, P = 1.7e-32, \( \chi^2 \) test) (Fig. 1D and Dataset S2). The G2, tightly located close to G1 (Fig. 1B), was identified as a subtype, which was represented by GATA3 mutations, while no significant fusion transcripts were detected (Fig. 1D). The G3, another subtype close to G1, contained patients all harboring SPII fusions (TCF7–SPI1 and STMN1–SPI1). Patients in G4, G5, and G6 exhibited overexpression of HOXA family genes (Fig. 1D). All patients in G4 harbored KMT2A fusions and MLLT10 rearrangements were regarded as key fusion events in G5, whereas HOXA10 fusions mainly occurred in G6 (Fig. 1D). T-ALLs in G7 and G8 were, respectively, characterized by TTX3 and TTX1 overexpression (Fig. 1D). The G9 subtype was unique because of the overexpression of NKX2-1 (interfering T cell differentiation by ectopic expression) (26). The G10 subtype was featured by diverse fusion events (such as STIL–TAL1, TAL2 fusions, LMO2 fusions, and LMO1 fusions), and also by the overexpression of TAL1 (participating in an oncogenic transcriptional program) (27) and LMO1 (altering T-cell differentiation together with TAL1) (28) in most patients (Fig. 1D).

In summary, transcriptome-driven molecular subtyping is biologically relevant, with each subtype associated with unique molecular abnormalities, namely: G1 (LYL1/LMO2 overexpression), G3 (SPII) fusion, G4 (KMT2A rearrangement), G5 (MLLT10 rearrangement), G6 (HOXA10 fusion), G7 (TTX3 overexpression probably involving TCR fusion), G8 (TTX1 overexpression probably involving TCR fusion), G9 (NKX2-1 overexpression), and G10 (TAL1/LMO1 overexpression). The identification of subtype G2 motivated further investigation into its molecular and functional mechanism.

Evidence for Subtypes from Nonsilent Gene Mutations in T-ALL. Using the previously established workflow (7, 9), we sought to explore evidence from our RNA-seq dataset, allowing the identification of nonsilent gene mutations with high sensitivity and specificity (SI Appendix, Fig. S1H). The number of nonsilent mutations detected in patients was significantly correlated with age (Spearman correlation coefficient \( R = 0.44, \ P < 0.0001 \)) (SI Appendix, Fig. S1I). We identified a total of 2,380 candidate mutated genes, with 78 recurring in >1% of T-ALL patients. These 78 genes were broadly grouped into 9 functional categories (C1 to C9): NOTCH signaling (C1), epigenetic regulators (C2), transcription factors (C3), PI3K-AKT-mTOR signaling (C4), JAK-STAT signaling (C5), RAS signaling pathway (C6), translation (C7), proliferation/apoptosis (C8), and others (C9) (Datasets S3–S5). Mutations distributed among 10 subtypes (G1 to G10) are detailed in SI Appendix, Fig. S2 and summarized based on functional categories (C1 to C9) (Fig. 2 A, Middle). The most frequently mutated genes included NOTCH1 (492 of 707, 69.6%), followed by FBXW7, PHF6, PTEN, and others (Fig. 2 A, Bottom). Most mutations in FBXW7, PHF6, and PTEN, three well-known tumor suppression genes, were loss-of-function in nature (29–31). Though with a much lower frequency (3.5%), GATA3 mutations were highly enriched (\( P = 8.7e-18 \), Fisher’s exact test) in G2 (Fig. 2A), thus supporting the notation of G2 (GATA3-mut) (Fig. 1D).

Mutated genes were functionally diverse (Fig. 2A), possibly acting as a coherent network. To support this, we integrated the knowledge of gene interactions (defined by Kyoto Encyclopedia of Genes and Genomes [KEGG] pathways). A network of mutated genes emerged, which was useful to explore relationships between mutated genes and subtypes (Fig. 2B). We next explored the percentage of mutations in each subtype and compared the frequency of the given mutations to that in other subtypes (Fig. 2C and Dataset S6). For example, mutations of NOTCH1 were found in all subtypes, but the rates were relatively lower in G1 (111 of 199, 55.78%), G4 (10 of 18, 55.56%), and G10 (183 of 278, 65.83%), suggestive of a tendency of independence, and higher in other subtypes, including G6 (11 of 11, 100%) and G2 (10 of 11, 90.9%), suggestive of cooccurrence (Fig. 2C). As expected, the GATA3 mutations most significantly clustered in G2 (Fig. 2C). For other subtypes, key findings are summarized: among patients in G1 (LYL1/LMO2), mutated genes in epigenetic regulators (PHF6, ASXL1, CHD4, EZH2, SETD2, DNMT3A, and HDJ2), transcription factors (WT1, RUNX1, ETV6, MED12, IKZF1), JAK-STAT signaling (JAK3, JAK1), RAS signaling (NRAS, KRA), and spliceosome complex (U2AF) (\( P < 0.05 \)) (Dataset S6) showed concurrent tendency (Fig. 2C). Gene mutations in the PI3K-AKT-mTOR signaling were enriched in G10 (TAL1/LMO1) (C4, \( P = 3.5e-24 \)) (Dataset S6). We also examined overall correlations between gene–gene (Dataset S6) and category–category correlations (Dataset S6). The results revealed mutations in PI3K-AKT-mTOR signaling (C4) were mutually exclusive with those in the JAK-STAT signaling (C5, \( P = 1.2e-5 \)) and the RAS signaling (C6, \( P = 1.5e-3 \)) (SI Appendix, Fig. S3 and Dataset S6). Of note, mutations in isocitrate dehydrogenases (IDH2) significantly cooccurred with the epigenetic modifier DNMT3A and NRAS mutations (SI Appendix, Fig. S3 and Dataset S6). We also noticed that RPL10(G88S) hotspot mutation formed two subclusters (G9 and G10) (SI Appendix, Fig. S2), which was reported to be associated with young age T-ALL, and altered T cell development by enhancing JAK-STAT signaling (32, 33).

GATA3 \( R^{276Q} \) Functions as a Driver Gene in T-ALL Leukemogenesis. Given that GATA3 point mutations were identified to signify subtype G2, we next addressed their functional features. GATA3 is a key transcriptional regulator for T cell development through binding to the DNA consensus sequence GATA (34). The elevation of GATA3 gene expression was found in T cell lymphoproliferative disorders (35, 36) and solid
tumors, such as breast cancer (SI Appendix, Fig. S4 A and B). Indeed, most mutations were clustered on the DNA binding domain with an intact open reading frame (ORF), in contrast to some cases of GATA3 mutations distributed in other subtypes with a truncated ORF of the genes (Fig. 3 A). Previously, the lack of or aberrant gene expression of GATA3 was linked to cancerogenesis, especially in leukemia (37). Despite limited sample size, the prognosis of G2 cases seemed to be poor (using available survival data, 5 dead and 2 alive in G2, 112 dead and 455 alive in other subtypes, $P = 4.8e-3$ by Fisher’s exact test).

The mutants identified in patients in G2 (GATA3-mut) were all located on the N-finger domain of GATA3 protein, such as R276Q ($n = 5$) (Fig. 3A). The expression level of GATA3 in G2 was significantly elevated, further supporting the importance of GATA3 in leukemogenesis (Fig. 3B). We also noticed that the expression levels of altered GATA3 were significantly higher in G2 (Fig. 3C), which was not observed in other subtypes (SI Appendix, Fig. S4 C and D). Although one case in G2 had lower GATA3 expression (SI Appendix, Fig. S3B), the variant allele frequency of GATA3 R276Q was 0.891 (Dataset S5) and this case harbored NRAS G13D.

To evaluate the binding affinity, we built two different GATA3/DNA complex sets (wrapping vs. bridging conformations), A B C 4o f1 0 https://doi.org/10.1073/pnas.2120787119 pnas.org

Fig. 2. The landscape of molecular interaction and pairwise relationship between nonsilent gene mutations. (A) Profiling of nonsilent gene mutations identified in 707 T-ALL RNA-seq. Mutation counts, gene mutations with high frequencies, and mutations in different categories are illustrated in three panels. In the Top, the number of mutations identified in RNA-seq data are illustrated as a barplot. In the Middle, genes with over 10% mutation frequency in T-ALL, as well as USP7 (9.5%) and GATA3 (used to discriminate GATA-mut subtype), are visualized. In the Bottom, mutation events in different categories are summarized using a blue label. (B, Left) Network visualization of mutated genes with edges defined by the knowledge of gene interactions from KEGG pathways. (Right) The same network but with nodes color-coded by the subtype-specific mutation frequency. (C) Comparison of the percentage of mutations in each subtype. Tendencies of cooccurrence and independence/exclusivity between gene mutations and subtypes are calculated, respectively. Red pies represent statistically significant cooccurrence, blue ones indicate statistically significant exclusivity, while gray ones show tendencies of gene mutation relationship that does not reach to statistical significance. Statistical significance of cooccurrence and exclusivity is calculated by comparing the mutations frequency in this subtype with other subtypes using $\chi^2$ test (when cases in all conditions $>5$) or Fisher’s exact test. Due to the limited sample sizes in some subtypes, some tendencies of relationship between gene mutations and subtypes could not always reach to statistical significance. Statistical results between mutations and subtypes are listed in Dataset S6.
Fig. 3. Schematic representation of GATA3 point mutations. (A) Protein structure of GATA3 and its mutations in T-ALL. All mutations on GATA3 are visualized on the upper area of the protein structure, and mutations in the N-finger domain identified in G2 are visualized in the lower area. (B) Boxplot of gene expression of GATA3 in each subtype. The dashed line represents the mean value of GATA3 in 707 T-ALLs. The P values are calculated by comparing with the mean gene expression of GATA3 using Wilcoxon rank-sum test. (C) Boxplot of count of GATA3 WT and altered GATA3 MUT reads in G2. P value is calculated using paired Wilcoxon rank-sum test. (D) Binding free energy (KJ/mol) reveals binding affinity of GATA3 (wild type, R276Q) protein and wrapping and bridging DNA sequence. (E) Volcano plot shows the differentially expressed genes between G2 (GATA3-mut) and GATA3 WT T-ALL. Each dot represents one gene. Genes significantly up-regulated in G2 (GATA3-mut) are colored in red, and significantly down-regulated in G2 (GATA3-mut) in blue. (F, L) Gene expression level of predicted bridging genes in T-ALL patients with different GATA3 genotypes; (Right) gene ontology results using up-regulated bridging genes in GATA3 R276Q cancer. (G) WISH results of rag1 RNA probes between GATA3 WT and GATA3 R276Q-mRNA injected embryos at 4 dpf. The phenotypes are defined as four groups: high, normal, mild, and extremely low according to the rag1 RNA expression in thymus. The percent is quantified (Right). P value is calculated using Fisher’s exact test. (H) WISH results of cmyb, α-globin, and γc RNA probes between GATA3 WT and GATA3 R276Q-injected embryos at 4 dpf. P value is calculated using Wilcoxon rank-sum test. (I) qRTP-PCR analysis of mRNA expression of the GATA3 downstream genes and rag1 in both GATA3 WT- and GATA3 R276Q-mRNA injected embryos at 4 dpf. The relative mRNA expressions are normalized to human GATA3. P values are calculated using Student’s t test.

Each containing the wild-type GATA3 (GATA3 WT) and mutant (GATA3 R276Q), and calculated the binding free energy to estimate the GATA3/DNA interaction (SI Appendix, SI Materials and Methods). Our model simulations showed that for the wrapping complex, the binding free energy was significantly higher in GATA3 R276Q/DNA (thus lower DNA-binding affinity) than in GATA3 WT/DNA, and no difference was observed for the bridging structure (Fig. 3D). To predict target genes that might be affected by wrapping/bridging motifs, we employed HOMER (39), combined with chromatin immunoprecipitation-sequencing data of GATA3 WT in human Jurkat cell lines (40) (Dataset S7), to define three sets of target genes: genes affected by wrapping motifs (wrapping targets), bridging motifs (bridging targets), and those insusceptible to wrapping/bridging motifs (other targets). Meanwhile, we identified 685 significantly up-regulated genes and 318 down-regulated genes in G2 (GATA3-mut), but not in GATA3 WT T-ALL cases falling into other subtypes (Fig. 3E and SI Appendix, Fig. S4G). These data supported the unique role of GATA3 in defining G2 as a distinct subtype. Among the 685 up-regulated genes in G2, 215 (31%) were GATA3 target genes, including 84 wrapping targets, 100 bridging targets, and 49 wrapping/bridging targets (Fig. 3F). These latter 49 genes were of functional relevance to negative regulation of myeloid cell differentiation (ZBTB16 and MEIS2), the CAMP signaling pathway (ATP1B1), and transcription factor binding (GATA3) (Fig. 3F). Notably, ZBTB16, also known as promyelocytic leukemia zinc finger PLZF, was reportedly to be involved in T cell lineage development (41) (SI Appendix, Fig. S4 H and I).

To examine the in vivo effect of GATA3 R276Q in hematopoiesis, we tried to use the zebrafish as a model. In this regard, we tested the leukemogenic role of the above-mentioned RPL16ΔC, an established leukemogenic mutation (32, 33), in a zebrafish experiment. Indeed, the results showed that cmyb, a
hematopoietic stem/progenitor cell (HSPC) marker, was aberrantly up-regulated in the caudal hematopoietic tissue of RPL10<sup>10885</sup> mRNA-injected embryos compared to RPL10<sup>10841</sup> as revealed by whole-mount in situ hybridization (WISH), reflecting an abnormal proliferation of HSPCs (SI Appendix, Fig. S5). Now that the system was feasible, we carried out the overexpression assay for both GATA3<sup>WT</sup> and GATA3<sup>R276Q</sup> in zebrafish. GATA3<sup>WT</sup> and GATA3<sup>R276Q</sup> mRNAs were injected into embryos for transient expression, followed by WISH examining the definitive hematopoiesis (Fig. 3 G and H). The expression of rag1 (a lymphocyte marker in zebrafish) was significantly up-regulated in GATA3<sup>R276Q</sup> mRNA-injected embryos in the thymus at 4 d postfecundation (dpf, P = 0.0076) (Fig. 3G), but no such altered expression was found with cmyb, actl1-globin, and lye (Fig. 3H), the latter two being markers of erythroid cells and neutrophils, respectively. qRT-PCR confirmed the up-regulation of GATA3 downstream genes (including gata3 (Fig. S6A)), Bcl2a/b, zbtb16a/b, meiz2a/bh, bel2a/bh, and spil1a and rag1 mRNA in the GATA3<sup>R276Q</sup> mRNA-injected group (Fig. 3I and Dataset S8). Taken together, our results confirmed GATA3<sup>R276Q</sup> as a driver for thymocyte proliferation in zebrafish embryos through either enhancing the effect of hematopoiesis-associated transcription factors (GATA3, ZBTB16, MEIS2) or activating target genes involved in T cell development pathways (including TGF-β, NOTCH, and Wnt/β-catenin signaling) (SI Appendix, Fig. S6A). These signaling cascades might collectively affect T cell proliferation and differentiation, eventually contributing to the pathogenesis of T-ALL.

**Association of Molecular Subtypes with T Cell Development Stages.** Given that genetic alterations distributed differently among subtypes, particularly ETP-ALL patients mainly found in G1 (LYL1/LMO2), we hypothesized that the subtypes might be inherently related to the maturation stages of T cells. In support of this, patients in G1 had low expression levels of T cell-related markers—such as CD1A, CD2, CD3E, CD4, and CD8A (immunophenotype-related genes)—but higher hematopoietic stem cell-related markers, such as CD34 (SI Appendix, Fig. S6B). Notably, the myeloid marker CD33 was highly expressed in G1 (SI Appendix, Fig. S6B). Furthermore, the expression patterns for the hematopoietic-related features in G2 and G3 were similar to those in G1, suggesting that T-lineage elements in G2 and G3 might be primitive (SI Appendix, Fig. S6B). Leukemic cells in G4 to G10 subtypes were likely at relatively late-stage T cell development, considering the decreased expression of HSPC-related markers and the increased expression of T cell-related markers (SI Appendix, Fig. S6B). To systematically associate subtypes with T cell development stages, we used the diffusion maps (42) for dimensionality reduction of 707 T-ALL, yielding three distinct branches (Fig. 4 A): branch 1 for patients mainly from G1 to G6, branch 2 (G7 and G8), and branch 3 (G9 and G10). These three branches differed in ETP status and T cell maturation stages, showing that ETP-ALL with pro/precortical immunophenotypes were enriched in branch 1, whereas postcortical and medullary patients were enriched in branch 3 (SI Appendix, Fig. S6C).

Next, we used available public data of T cell expression functional clusters (43) (Dataset S9) to characterize subtypes and branches. The T cell differentiation stages in branch 1 patients were the earliest, branch 2 patients at an intermediate stage, and branch 3 patients at the late stage (the mature T-ALL stage) (Fig. 4B). Together with dysregulated leukemogenic factors and subtypes, T-ALL patients thus could be divided into three differentiation arrest branches: LYL1/LMO2/SPI1/HOXA high expression (branch 1, G1 to G6), TLX3/TLXI high expression (branch 2, G7 to G8), and NKX2-1/TAL1/LMO1 high expression (branch 3, G9 to G10). By using both ETP signature (ETP status) and precortical signature (T cell development status), we generated the signature-specific enrichment score for each patient (Fig. 4C and Dataset S10). Patients with LYL1/ LMO2/SPI1/HOXA tended to have higher scores for both ETP and precortical signatures, whereas patients with NKX2-1/ TAL1/LMO1 had lower scores for both signatures (Fig. 4G).

We also compared the age and gender composition between the three branches and found that the LYL1/LMO2/SPI1/ HOXA branch had the highest percentage of adult patients (P = 1.3e-26, χ<sup>2</sup> test) (Fig. 4D) and harbored more mutations (Fig. 4E). Functional categories were distributed differently among the three branches: mutations in NOTCH signaling (C1) were enriched in the TLX3/TLXI patients; mutations in epigenetic regulators (C2), transcription factors (C3), JAK-STAT signaling (C5), RAS signaling (C6), and proliferation/apoptosis (C8) were in the LYL1/LMO2/SPI1/HOXA and TLX3/TLXI; and mutations in PI3K-AKT-mTOR signaling (C4) were specifically concentrated in the NKX2-1/TAL1/LMO1 branch (Fig. 4F).

In light of the prevailing perspectives (1, 5) and the knowledge obtained in this study, a working model of T-ALL leukemogenesis was proposed, with four key points (Fig. 4G): 1) the accumulation of genetic abnormalities, such as gene fusions and cancer driver mutations, could cause the dysregulation of different leukemogenic factors, ultimately blocking normal T cell development; 2) for T-ALL patients with the LYL1/LMO2/ SPI1/HOXA (G1 to G6), the immunophenotype of blasts might represent the population blocked at a near HSC or very early T cell development stage (i.e., ETP-ALL or near ETP, pro/precortical or cortical), and the age-onset for leukemogenesis tended to be higher (with over 50% adult patients); 3) for T-ALL patients with the TLX3/TLXI (G7 and G8), leukemic blasts could be blocked at an intermediate stage, with the cortical or postcortical immunophenotype of T-ALL cells; and 4) the immunophenotype of T-ALL patients with the NKX2-1/ TAL1/LMO1 (G9 and G10) might correspond to cortical, postcortical, or mature T cell counterparts at the late stage.

**Exploring Genomic, Expression, and Cellular Correlates likely Explaining Differences in Adult and Pediatric T-ALL Patients.** Correlating age with mutations in T-ALL, we found that DNMT3A and IDH2 mutations tended to occur in the relatively elderly patients, with the mean age of 53 y for IDH2 and 48 y for DNMT3A (Fig. 5A and SI Appendix, Fig. S7A). When the G1 group, which contained the majority of ETP-ALL in this series, was further scrutinized, the mutations rates of DNMT3A and IDH2, two genes with high lesion frequencies in acute myeloid leukemia (AML), were concentrated in about 44% of the G1 cases with the age of over 40 y (SI Appendix, Fig. S7B). The mutations tended to be frameshift and stop-gain for DNMT3A, suggesting loss-of-function (SI Appendix, Fig. S7C) (44), whereas hotspot missense mutations R140Q were observed in IDH2, in agreement with a gain-of-function alteration (SI Appendix, Fig. S7D) (45). It was reported that DNMT3A and IDH2 mutations could cooperate to induce AML (44), but such cooperation was not yet identified in T-ALL. Mutated genes in epigenetic regulators (IDH2, DNMT3A, CHD4, ASXL1, CREEB, and EZH2), transcription factors (IKZF1, ETV6, and RUNX1), JAK-STAT signaling (JAK3 and JAK1), and RAS signaling (NRAS) were enriched in
**Fig. 4.** Dimensionality reduction analysis revealing T cell development in different subtypes. (A) Visualization of the dimensions calculated by diffusion map using 707 T-ALL patients. The top 5% variance genes in RNA-seq data are subjected to diffusion map analysis and the first three diffusion components are visualized using three-dimensional plots. Each point represents one sample. (B) Gene-expression patterns of signatures of different functional clusters. These clusters were differentially expressed in different T cell stages. The Left heatmap shows the expression levels of functional clusters in different T cell stages, the Middle heatmap shows the expression levels in different subtypes, while the Right heatmap shows the expression levels in different branches. Expression is calculated using the mean value of the genes and then scaled as the row z-score. (C) Scatter and density plot of enrichment score (ES) for ETP status. (D) Bar plot of the percentage of patients according to age (Upper) and gender (Lower) in each dysregulated leukemogenic factor branch. P values are calculated using Wilcoxon rank-sum test. (E) Model of the association between the accumulation of genetic abnormalities, the dysregulation of leukemogenic factors, T cell stages, and age in T-ALL leukemogenesis.
adult T-ALL, while those in \textit{FBXW7}, \textit{BCL11B}, and \textit{RPL10} were more likely to occur in pediatric T-ALL ($P < 0.05$) (Fig. 5A). Regarding gene fusions, we sorted the patients by their age at diagnosis (Fig. 5B). \textit{SET-NUP214}, \textit{NUP98} fusions and \textit{ZFP36L2} fusions were significantly enriched in adult patients, whereas \textit{SPI1} fusions, \textit{NKX2-1} fusions, and \textit{STIL-TAL1} were more likely to occur in pediatric patients ($P < 0.05$) (Fig. 5B). Since differences in genetic abnormalities/expression did exist between pediatric and adult patients, we searched for genes that were significantly correlated with age, especially those in pathways of vulnerability for potential therapeutic targets (Fig. 5B). Indeed, the expression levels of two therapeutic targets, \textit{BCL2} (participating in apoptosis) and \textit{LCK} (involved in preTCR activation) (46), were found oppositely correlated with age in T-ALL (SI Appendix, Fig. S7E). Other targets (\textit{JAK1}, \textit{ABLI}, and \textit{FLT3}) were mainly up-regulated in adult T-ALL (Fig. 5B). Of note, \textit{FLT3} and its related gene signatures (such as the \textit{AML} pathway and \textit{BCL2}) was highly expressed in G1, G2, G3, and G7, especially in ETP and pro cortical T-ALLs (SI Appendix, Fig. S8 A–Q), indicating \textit{FLT3} inhibitors as a potential therapeutic target for future study.

It was recently established that ETP-ALL share features with mixed phenotype acute leukemia (MPAL) (47). We thus used AML gene signatures for enrichment analysis in an ETP-ALL subset and found a significantly augmented AML enrichment score with age (SI Appendix, Fig. S8B). Applying the xCell algorithm (48) and the 17-gene stemness score (49), we were able to infer the proportion of hematopoietic stem cell, common lymphoid progenitor, and common myeloid progenitor, and show a much higher likelihood of adult patients than pediatric ones to exhibit MPAL (SI Appendix, Fig. S8 D and E). Extending the leukemogenesis model of MPAL, we deduced that genetic lesions might occur at an earlier HSPC stage in adult T-ALL, resulting in abnormal proliferation/differentiation/
apoptosis of myeloid and lymphoid lineages, which might be significant in explaining the difference in prognosis and treatment responses observed between pediatric and adult T-ALL (SI Appendix, Fig. S8 F and G, respectively illustrating pediatric and adult leukemic hematopoiesis models).

Discussion

The transcriptomic landscape of 707 T-ALLs unveiled 10 subtypes in this study, which was superior to previous classifications of T-ALL that depended on the single/double combination of dysregulated leukemogenic factors, such as LOM2, LYL1, HOXA family genes, TLX3/1, NXX2-1, LMO1, and TAL1/2. Limited by the small sample size, we previously classified patients into three parts, namely TLX1/3/HOXA, ETP/LYL1/HOXA, TAL1/LMO1 (9). The classification of ALL based on gene-expression patterns in larger ALL cohorts could help to identify rare subtypes (7). In the present study, GATA3 mutations (G2) were first clarified as a subtype of T-ALL with a poor prognosis. Because previous studies have reported GATA3 in solid tumors, including breast cancer (35, 36), how the role of GATA3 dysregulation participated in T cell commitment would be further investigated in T-ALL and other T cell disorders. Additionally, three subtypes with elevated expression of HOXA family genes were revealed—namely G4 (KMT2A+), G5 (MLLT10+), and G6 (HOXA10-fus)—each representing a small number of T-ALL patients, similar to G2, which might be the reason why they were not found as independent subtypes in a previous study (9).

Meanwhile, the expanded sample size of adult T-ALL allowed us to conduct a comparison of abnormal genome/transcriptome landscapes between pediatric and adult patients (Fig. 6). We illustrated that some fusions/mutations differed significantly in frequencies between the two age groups. Adult patients tended to harbor more nonsilent mutations than pediatric ones, especially in epigenetic regulators—with DNMT3A and IDH2 being the most representative ones—JAK-STAT signaling, and RAS signaling pathways. These sequence mutations could cooperate with aberrant gene expression and exert an effect on clinical outcomes. Our results provide evidence that the more complex genetic abnormalities in leukemic cells in adults than in children may contribute to the unfavorable prognosis in the former age group. In addition, adult patients, particularly those aged over 40 y, are more likely to bear the features of MPAL, which render the malignant cells less sensitive to the current therapeutic agents. In this regard, it may be interesting to note the emergence of some potential therapeutic drug targets in adult T-ALL, such as BCL-2 and FLT3.

There are some limitations to note in our study. First, the number of patients in some subtypes is still limited. Second, using RNA-seq alone to identify copy number variations or deletions, such as of CDKN2A/CDKN2B, is challenging. Fusions involving TLX1/TLX3 and LMO1/LMO2 are hardly identified from RNA-seq data, although their high expression levels are suggestive of the existence of fusion to TCR. Third, without samples from normal tissues as control and genomic DNA sequencing data in most cohorts, nonsilent mutations were inferred based on the previously reported mutation profiles in T-ALL and other leukemia (5, 7, 9, 16–21) and improvement of mutation-calling pipelines (described in SI Appendix, SI Materials and Methods).

Despite these limitations, this work can facilitate an in-depth understanding of the biological nature of T-ALL. The dimensionality reduction analysis proves useful to determine the associations between molecular subtypes and phenotypes according to the stages of blockage in T cell development.

In conclusion, we identified 10 subtypes of T-ALL, characterized their genetic alteration patterns, and investigated the associations between these subtypes and T cell development stages. These results revealed that the involvement of T cell differentiation stage was earliest in G1 and latest in G10. Based on the dysregulated leukemogenic factors, we have revealed relative mutation/abnormal expression features in adult and pediatric T-ALL patients. Furthermore, our study lends support to the feasibility of RNA-seq as a clinical platform for the classification of T-ALL.

Materials and Methods

Patients. Patients in cohorts 1 to 6 were obtained from public database. Patients in cohort 7 were from the Hematological Biobank, Jiangsu Biobank of Clinical Resources during 2016 to 2019. Patients in cohort 8 were from a multicenter study under the coordination of the Shanghai Institute of Hematology, including Chinese People’s Liberation Army General Hospital and the First Affiliated Hospital, Zhejiang University College of Medicine, and Second Hospital of Dalian Medical University, these cohorts being followed from 2016 to 2020. Informed consent for cohorts 7 and 8 patients in the study was obtained by the participating centers. This research was approved by Ruijin Hospital Ethics Committee. Detailed information of the patients is listed in SI Appendix, SI Materials and Methods and Dataset S1.

RNA-seq Analysis, Mutations and Fusions Calling, Zebrafish Experiment, and T Cell Differentiation Stage Analysis. Detailed materials and methods are provided in SI Appendix, SI Materials and Methods. All animal experiments were approved by the Committee of Animal Use for Research at Shanghai Jiao Tong University School of Medicine (China).

Data Availability. RNA sequencing data generated in this study are deposited at the National Omics Data Encyclopedia (NODE) (accession code OEP002748). Previously published data used for this work were Liu et al. (5), Seki et al. (17), Qian et al. (16), Yasuda et al. (18), Chen et al. (9), and Autry et al. (19).

ACKNOWLEDGMENTS. We thank the Center for High Performance Computing of Shanghai Jiao Tong University for providing computing support; and Prof. Jinghui Zhang and Prof. Jun J. Yang from the Department of Pharmaceutical
Sciences, St. Jude Children's Research Hospital, for their helpful advice on data analyses. This work was supported by the National Natural Science Foundation of China General Program (18770205, 32170663, 81670147, 8116148030), Antrag M-0377; the State Key Laboratory of Medical Genomics; the Double First-Class Project (WF510162602) from the Ministry of Education; the Shanghai Collaborative Innovation Program on Regenerative Medicine and Stem Cell Research (2019CX001); the Overseas Expertise Introduction Project for Discipline Innovation (111 Project; B17029); the Shanghai Shenkang Hospital Development Center (SHDC2020CR5002); the Shanghai Major Project for Clinical Medicine (2017Z201002); and the Innovative Research Team of High-level Local Universities in Shanghai.

Author affiliations: 1Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 200025 Shanghai, China; 2State Key Laboratory of Biochip Technology, Shanghai Jiao Tong University, 200240 Shanghai, China; 3State Key Laboratory of Microbial metabolism, joint International Research Laboratory of Metabolic & Developmental Sciences, Department of Bioinformatics and Biostatistics, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 200240 Shanghai, China; 4Department of Hematology, National Clinical Research Center for Hematologic Diseases, Jiangsu Institute of Hematology, The First Affiliated Hospital of Nanjing Medical University, 200002 Nanjing, China; 5School of Medicine, University of Singapore, 117599 Singapore; 6Institute of Biostatistics and Biomathematics, The Chinese University of Hong Kong, 999077 Hong Kong, China; 7Institute of Hematological Research, Shanghai Jiao Tong University, 200240 Shanghai, China; 8Department of Clinical Pharmacy, School of Pharmaceutical Sciences, Soochow University, 215006 Suzhou, People’s Republic of China; 9Department of Clinical Medicine, Soochow University, 215006 Suzhou, People’s Republic of China; 10Division of Hematology and Rheumatology, Shanghai University of Traditional Chinese Medicine, Faculty of Medicine, 577852 Osaka, Japan; 11Department of Hematology, Atomic Bomb Disease Institute, Nagasaki University, 8528652 Nagasaki, Japan; 12Clinical Research Center, Nagoya Medical College, Showa Hospital, 460001 Nagoya, Japan; 13Center for Translational Research in Acute Leukaemia, Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, 117597 Singapore; 14Viva-Univers Children’s Cancer Center, Khoa Teck Puat-National University Children’s Medical Institute, National University Hospital, 119074 Singapore; 15Cancer Science Institute of Singapore, National University of Singapore, 117599 Singapore; 16Department of Hematology and Oncology, Nagoya University Graduate School of Medicine, 4686850 Nagoya, Japan; 17Department of Integrated Health Sciences, Division of Cellular and Genetic Sciences, Nagoya University Graduate School of Medicine, 4610047 Nagoya, Japan; 18Center for Biomolecular Big Data, The First Affiliated Hospital, School of Medicine, Zhejiang University College, 310003 Hangzhou, China; 19Clinical Center, Zhejiang University, 310058 Hangzhou, China; 20University of Hong Kong, Department of Medicine (2017ZZ01002); and the Innovative Research Team of High-level Local Universities in Shanghai.

Author contributions: X.J.S., J.-Q.M., Z.C., J.-Y.H., and S.-J.C. designed research; F.Z., H.F., G.L., L.J., B.C., D.D.M., Y.-F.L., J.W., L.J.P., C.F., H.-F.C., J.-X.M., T.A.M., S.-Y.W., X.-J.S., J.-Q.M., Z.C., J.-Y.H., and S.-J.C. performed research; J.-F.L., Q.-L.Z., H.W., H.A., J.-H.W., Y.-J.L.N, L.C., Z.S., I.M., Y.M., T.-P.D., X.Y., J.-S.Y., A.E.J.Y., D.-P.W., H.K., F.H., and J.J. contributed new reagents/analytic tools; Y.-F.L., J.W., L.J.P., H.W., H.A., T.A.M., J.-H.W., Y.-J.L.N, S.-N.C., Q.W., H.L., Z.S., Z.M., Y.M., T.Y., L.P.D., X.-J.Y., J.-S.Y., A.E.J.Y., D.-P.W., H.K., F.H., J.J., and Q.-M. contributed data and provided critical information; Y.-T.D., H.F., and J.-Y.H. analyzed data; and Y.-T.D., Z.-C., J.-Y.H., and S.-J.C. wrote the paper.

1. T. Girardi, C. Viccis, K. DeKremer, The genetics and molecular biology of TALL. Blood 129, 1113-1123 (2017).
2. C. H. Pu, M. V. Belling, J. R. Downing, Acute lymphoblastic leukemia. N. Engl. J. Med. 350, 1535-1548 (2004).
3. S. Chiaretti, R. Fea, T-cell acute lymphoblastic leukemia. Haematologica 94, 168-162 (2009).
4. C. H. Pu, W. E. Evans, Treatment of acute lymphoblastic leukemia. N. Engl. J. Med. 354, 166-178 (2006).
5. Y. Liu et al., The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. Nat. Genet. 49, 1211-1218 (2017).
6. I. Iarussi, C. G. Mullighan, Genetic basis of acute lymphoblastic leukemia. J. Clin. Oncol. 37, 975-983 (2019).
7. J. F. Li et al., Transcriptional landscape of B cell precursors acute lymphoblastic leukemia based on an international study of 1,223 cases. Proc. Natl. Acad. Sci. U.S.A. 115, E11711-E11720 (2018).
8. Y. Liu et al., Genomic profiling of adult and pediatric B-cell acute lymphoblastic leukemia. Blood 137, e140-156 (2016).
9. J. B. Jeron et al., Identification of fusion genes and characterization of transcriptome features in t-cell acute lymphoblastic leukemia. Proc. Natl. Acad. Sci. U.S.A. 115, 373-378 (2018).
10. L. Belver, A. Fernando, The genetics and mechanisms of T-cell acute lymphoblastic leukemia. Nat. Rev. Cancer 16, 494-506 (2016).
11. Q. Chen et al., Coding sequences of the tal-1 gene are disrupted by chromosome translocation in human T-cell leukemia. J. Exp. Med. 172, 1403-1408 (1990).
12. P. Van Wervenbergen et al., the recurrent S100P2/7 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia. Blood 111, 4684-4690 (2008).
13. E. P. Norondo et al., Brazilian Collaborative Study Group of Acute Leukemia, The profile of immunophenotype and genotype alterations in subsets of pediatric T-cell acute lymphoblastic leukemia. Front. Oncol. 9, 316 (2019).
14. E. Cauttan-Smith et al., Early T-precursor leukemia: A subtype of very high-risk acute lymphoblastic leukemia. Lancet Oncol. 10, 147-156 (2009).
15. U. Koch, F. Radke, Mechanisms of T-cell development and transformation. Annu. Rev. Cell Dev. Biol. 27, 539-562 (2011).
16. M. Qian et al., Whole-transcriptome sequencing identifies a distinct subtype of acute lymphoblastic leukemia with predominant genetic abnormalities of EP300 and CREBBP. Genome Res. 27, 185-195 (2017).
17. M. Seki et al., Recurrent SPT1 (PU.1) fusions in high-risk pediatric T-cell acute lymphoblastic leukemia. Nat. Genet. 49, 1274-1281 (2017).
18. T. Yauda et al., Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. Nat. Genet. 48, 569-574 (2016).
19. R. J. Autley et al., Integrative genomic analyses reveal mechanisms of glucocorticoid resistance in acute lymphoblastic leukemia. Nat. Genet. 49, 1274-1281 (2017).
20. L. Jiang et al., Multidimensional study of the heterogeneity of leukemia cells in T-B, acute myelogenous leukemia identifies the subtype with poor outcome. Proc. Natl. Acad. Sci. U.S.A. 117, 20117-20126 (2020).
21. J. Kong et al., Genetic and transcriptomic characterization of natural killer T leukemia. Cancer Cell 37, 403-419 e6 (2020).
22. Y. Zhong, L. Jiang, H. Hiai, S. Toyokuni, Y. Yamada, Overexpression of a transcription factor LYL1 induces T- and B-cell lymphoma in mice. Oncogene 26, 6937-6943 (2007).
23. D. J. Curtis, M. P. Armand, The molecular basis of Lmo2-induced T-cell acute lymphoblastic leukemia. Clin. Cancer Res. 16, 5618-5623 (2010).
24. A. Champushekar et al., Regulation of early T-lineage gene expression and developmental progression by the progenitor cell transcription factor PU.1. Genes Dev. 29, 832-848 (2015).