SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity

Maxime Caron[1,3], Pascal St-Onge[3], Thomas Sontag[3], Yu Chang Wang[2], Chantal Richer[3], Ioannis Ragoussis[1,2], Daniel Sinnett[3,4] ✉ & Guillaume Bourque[1,2,5] ✉

Childhood acute lymphoblastic leukemia (cALL) is the most common pediatric cancer. It is characterized by bone marrow lymphoid precursors that acquire genetic alterations, resulting in disrupted maturation and uncontrollable proliferation. More than a dozen molecular subtypes of variable severity can be used to classify cALL cases. Modern therapy protocols currently cure 85–90% of cases, but other patients are refractory or will relapse and eventually succumb to their disease. To better understand intratumor heterogeneity in cALL patients, we investigated the nature and extent of transcriptional heterogeneity at the cellular level by sequencing the transcriptomes of 39,375 individual cells in eight patients (six B-ALL and two T-ALL) and three healthy pediatric controls. We observed intra-individual transcriptional clusters in five out of the eight patients. Using pseudotime maturation trajectories of healthy B and T cells, we obtained the predicted developmental state of each leukemia cell and observed distribution shifts within patients. We showed that the predicted developmental states of these cancer cells are inversely correlated with ribosomal protein expression levels, which could be a common contributor to intra-individual heterogeneity in cALL patients.

Childhood acute lymphoblastic leukemia (cALL) is the most frequent pediatric cancer, accounting for ~25% of all pediatric tumors. This cancer is characterized by bone marrow lymphoid precursors that acquire sequential genetic alterations, resulting in disrupted maturation and uncontrollable proliferation. Precursor B cell ALL (B-ALL) represent ~85% of cases and precursor T cell ALL (T-ALL) ~15%, which can be further subdivided into more than a dozen molecular subtypes. The high hyper diploid cases (HHD) and those harboring the t(12;21) [*ETV6/RUNX1*] rearrangement represent about ~60% of B-ALL cases and are associated with a good prognosis[1,2]. Other less frequent (<10%) subtypes, such as KMT2A-rearranged, t(9;22) [*BCR/ABL1*] or T-ALL are associated with less favorable outcomes[1,2]. Five-year event-free survival rates have seen remarkable progress since the 1960s and are now up to ~85–90% thanks to patient risk stratification and effective treatment combinations[3]. Despite this therapeutic success refractory and relapsing cALL still represents one of the most frequent cause of death by disease in pediatrics[4]. This disease is primarily caused by initiating genetic lesions in lymphoid progenitors (chromosomal abnormalities) followed by secondary genetic lesions (copy number alterations, point mutations) leading to leukemogenesis[3,5].

Somatic mutations at diagnosis arise in many genes involved in multiple biological pathways such as B cell development (e.g. *PAX5, IZKF1*), signaling (e.g. *RAS, NRAS, CRLF2*) epigenetic regulation (e.g. *CREBBP*,

[1]Department of Human Genetics, McGill University, Montréal, Québec, Canada. [2]McGill Genome Centre, Montréal, Québec, Canada. [3]CHU Sainte-Justine Research Center, Montréal, Québec, Canada. [4]Department of Pediatrics, University of Montreal, Montréal, Québec, Canada. [5]Canadian Center for Computational Genomics, Montréal, Québec, Canada. ✉e-mail: daniel.sinnett@umontreal.ca; guil.bourque@mcgill.ca

1

Sample	Number of cells	Mean Reads per Cell	Median Genes per Cell	Number of Reads	Sequencing Saturation	Reads Mapped Confidently to Genome	Fraction Reads in Cells	Total Genes Detected	Median UMI per Cell
ETV6-RUNX1_1	2,776	76,003	1,830	210,987,037	84.0%	92.0%	89.8%	18,397	5,312
ETV6-RUNX1_2	6,274	52,662	1,239	330,404,706	87.9%	90.5%	85.1%	20,005	2,435
ETV6-RUNX1_3	3,862	85,994	1,181	332,110,056	90.3%	85.3%	80.8%	19,909	2,815
ETV6-RUNX1_4	5,069	59,411	1,304	301,155,525	84.3%	83.5%	88.3%	19,414	3,940
HHD_1	3,728	83,388	1,579	310,871,911	78.3%	89.1%	77.1%	19,061	4,712
HHD_2	5,013	64,741	2,059	324,551,184	73.4%	87.1%	93.3%	20,798	5,643
PRE-T_1	2,959	109,972	2,114	325,408,071	80.5%	85.3%	74.9%	19,775	6,309
PRE-T_2	2,748	118,053	1,685	324,412,280	86.1%	86.7%	81.8%	20,568	5,037
PBMMC_1	1,612	303,228	880	488,803,883	NA	NA	78.5%	NA	2,573
PBMMC_2	3,105	101,124	1,091	313,990,607	81.0%	79.2%	79.9%	19,927	6,408
PBMMC_3	2,229	145,472	1,085	324,258,554	91.4%	87.1%	61.6%	18,793	3,008

Table 1. Metadata of samples used in the study.

KMT2D) and cell cycling (e.g. *CDKN2A*)[6]. This led to the proposition of clonal evolution characterized by a branching model in which leukemic subclones acquire sequential genetic alterations conveying fitness advantages[6]. In a significant number of relapse cases, ancestral minor subclones present at diagnosis become resistant and positively selected under treatment[6,7]. Next-generation sequencing (NGS) has recently provided meaningful insights toward the characterization of somatic events providing selective advantages to tumor cells. Mutations in RAS pathway genes have been associated with early relapse and chemoresistance in cALL whereas mutations in the cytosolic 5′-nucleotidase II gene (*NT5C2*) were shown to be involved in resistance to treatment[8,9]. A study on the subclonal nature of childhood leukemia revealed that the majority of patients harbor at least one genetic subclone at diagnosis and three quarters of relapse cases were from resistant minor ancestral subclones[6]. Another study found the presence of at least two distinct genetic subclones in five out of six patients using targeted single-cell DNA sequencing[10]. In this context, we reported different clonal dynamics between early (during treatment) and late (>36 months after diagnostic) bone marrow relapse events and showed variable clonal origin in late relapse events[11].

Genetic alterations of subclones can provide information on the perturbed functions of genes and their corresponding pathways but the transcriptional impacts of these alterations in cALL are unknown as intra-individual transcriptional heterogeneity is not well characterized. Additionally, subclonal epigenetic modifications could cause changes in gene expression that cannot be captured by looking at genetic alterations only. Transcriptional changes at the single cell level can be assessed using recent single cell technologies that enable the transcriptional profiling of thousands of single cells, including droplet-based whole transcriptome profiling (scRNA-seq)[12]. The prognostic value of scRNA-seq was previously demonstrated in a single cell gene expression study that assessed the transcriptional heterogeneity of patients initially assigned to a unique subtype that were further classified into additional subgroups with distinct survival[13]. In this study we sought to uncover the extent and nature of transcriptional heterogeneity in cALL patients at diagnosis, identify distinct subpopulations of cancer cells and uncover deregulated genes and pathways. The role of other cellular processes beyond the cell cycle that can lead to intratumoral transcriptional heterogeneity are not well characterized. These intra-individual differences could provide information on the impact of genetic or epigenetic alterations on the expression programs of these cells and help us gain additional knowledge on the mechanisms of disease development.

Results

We generated single cell gene expression data from 39,375 pediatric bone marrow mononuclear cells (BMMCs) from eight cALL patients of common subtypes (4 *ETV6-RUNX1*, 2 High Hyper Diploid and 2 T-ALL) having >50% blasts and 3 healthy donors (Tables 1 and 2).

We assessed the quality of our data in healthy pediatric BMMCs with available data in healthy adult BMMCs (10X Genomics, Methods). We recovered the expected cell types using known cell surface marker genes (Fig. 1A–D).

In healthy cells, the predicted cell cycle phases showed a higher proportion of cycling cells in B cells and immature hematopoietic than in other cell types (Fig. 1E). By combining healthy pediatric BMMCs with cALL cells (n = 38,922 after quality control), we observed distinct clusters of healthy (PBMMCs) and cancer cells (Fig. 2A).

Between 2 and 60% of cALL cells per sample clustered with healthy pediatric BMMCs of different cell types (Fig. 2C, Supplementary Fig. 1). These cells likely represent non-cancerous cells normally found in samples of variable tumor purity (due to disease severity or technical variability), rather than lineage switching cancer cells or cancer cells having healthy-like transcriptional profiles, which is supported by copy number profiles that are similar to those of PBMMCs (Supplementary Fig. 2). When looking at the predicted cell cycle phases of cALL cells, we observed a continuous spectrum of phases G1 → S → G2/M on the UMAP representation (Fig. 2B). For six out of eight patients, cells were mostly in the G1 phase (Fig. 2D). Many methods can correct for different sources of transcriptional variation[14,15], however regressing out the cell cycle phase in our data failed to completely remove this effect as we could still observe clusters of cells in cycling phases on UMAP (Supplementary Fig. 1). Thus, in further analysis, we decided to reduce the expression variability by keeping cancer cells that did not cluster with healthy cells (remaining n = 25,788; ~79.5%) and that were in G1 phase only (remaining n = 16,731; ~51.6%; Fig. 3A).

Mean Reads per Cell	Median Genes per Cell	Number of Reads	Sequencing Saturation	Reads Mapped Confidently to Genome	Fraction Reads in Cells	Total Genes Detected	Median UMI per Cell
76,003	1,830	210,987,037	84.0%	92.0%	89.8%	18,397	5,312
52,662	1,239	330,404,706	87.9%	90.5%	85.1%	20,005	2,435
85,994	1,181	332,110,056	90.3%	85.3%	80.8%	19,909	2,815
59,411	1,304	301,155,525	84.3%	83.5%	88.3%	19,414	3,940
83,388	1,579	310,871,911	78.3%	89.1%	77.1%	19,061	4,712
64,741	2,059	324,551,184	73.4%	87.1%	93.3%	20,798	5,643
109,972	2,114	325,408,071	80.5%	85.3%	74.9%	19,775	6,309
118,053	1,685	324,412,280	86.1%	86.7%	81.8%	20,568	5,037
303,228	880	488,803,883	NA	NA	78.5%	NA	2,573
101,124	1,091	313,990,607	81.0%	79.2%	79.9%	19,927	6,408
145,472	1,085	324,258,554	91.4%	87.1%	61.6%	18,793	3,008

Table 2. Single cell statistics of samples used in the study. *Pediatric control PBMMC1 was generated from two independent libraries to increase the total number of cells and sequencing runs were aggregated using Cell Ranger.

We looked at the transcriptional profiles of these cancer cells using non-supervised hierarchical clustering of the hundred most variable genes. We observed two distinct clusters of B-ALL and T-ALL cells as shown by the expression of the *CD79A/B* and *CD3D* surface markers (Fig. 2E, Supplementary Figure 1). We also noticed transcriptional differences between B-ALL subtypes (e.g. *RAG1* over expressed in *ETV6-RUNX1* samples) and individuals (e.g. *TCL1B* over expressed in samples ETV6.RUNX1.3/ETV6.RUNX1.4 and *CD1E* over expressed in sample PRE-T.1). These transcriptional profiles corroborate expected cell surface marker expression for these types of cells and indicate good quality of the single cell expression data.

Intra-individual transcriptional heterogeneity. We assessed the intra-individual transcriptional heterogeneity by identifying expression profile signatures unique to subsets of cancer cells within a sample. These intra-individual transcriptional differences could highlight deregulated genes in potentially resistant subclones or reveal epigenetic changes driven by subclonal genetic alterations. We applied a strategy that would return the most representative clustering solution over a range of clustering resolutions. Multiple clustering solutions were generated over a range of parameters and pairwise Adjusted Rand Index (ARI) values were computed (Fig. 3B). We retained the clustering solution that had the highest mean ARI and identified at least two clusters in six samples, but no intra-individual transcriptional clusters for samples ETV6.RUNX1.4 and PRE-T.2 (Fig. 3C). Transcriptional differences between subsets of cell within a sample require enough cells (>100) to be properly powered for differential expression analyses[16]. Thus, we discarded clusters with less than 10% of total cells for samples ETV6.RUNX1.3 and HHD.2 (remaining n = 16,162). This filtering steps resulted in 5 samples with two transcriptional clusters with proportions of cells ranging from ~25 to 50% (Fig. 3D). Repeating the same analysis using cells in S + G2/M phases (n = 9417) identified clusters that were similar to those identified in G1, however differences were observed for some samples which could be due to the lower number of cells in the S + G2/M phases, cell cycle heterogeneity and clustering resolution (Supplementary Figure 3). Identifying transcriptional clusters in each sample individually returned similar results for most samples (Supplementary Figure 4). We then performed intra-individual differential gene expression analyses for the two subsets of cells within these 5 samples. Genes related to B and T cell maturation (e.g. *CD37, CD79B, JCHAIN, IGLL1, VPREB3, CD52*), ribosomal protein genes (*RPS-*, RPL-**) and cancer/stress genes (e.g. *PRAME, ARHGDIB, FOS, JUN*) (Fig. 3E, Supplementary Table 1) were within the most significantly deregulated genes between clusters in those samples. Gene ontology analyses on the five differential expression gene lists led to a more general understanding of the modulated biological pathways within each sample (Supplementary Table 2). The samples were then clustered based on the top ten most significantly enriched biological pathways. We observed two distinct clusters: one cluster of *ETV6-RUNX1* samples showing modulated pathways in B cell activation/differentiation, cell proliferation/death and regulation of expression and metabolic processes and another cluster of HHD and pre-T samples showing modulated pathways in translation initiation and protein synthesis (Fig. 3F). These results demonstrate that we can detect distinct sources of transcriptional variation between subsets of cells within a sample.

Genetic alterations and transcriptional variability. To uncover genetic alterations that could be linked to intra-individual transcriptional heterogeneity, we first looked whether large cluster specific copy number variants (CNVs) could be linked to transcriptional clusters. We generated thirty equal size 'metacells' for each cluster to reduce noise and called copy number events using healthy pediatric BMMCs as controls. We recovered most copy number events seen with exome sequencing data and identified cluster specific copy number events (Fig. 4A, Supplementary Figure 5).

We observed cluster specific copy number gains on chromosomes 21 and 22 in sample ETV6.RUNX1.1 and deletions of chromosome 5 and the p arm of chromosome 12 in sample ETV6.RUNX1.2. The chr22 gain and chr5 deletion were found to be subclonal when looking at CNV profiles of individual cells, but we could not confirm the other cluster specific alterations, which could be due to the inherent noise of calling CNVs using single cells
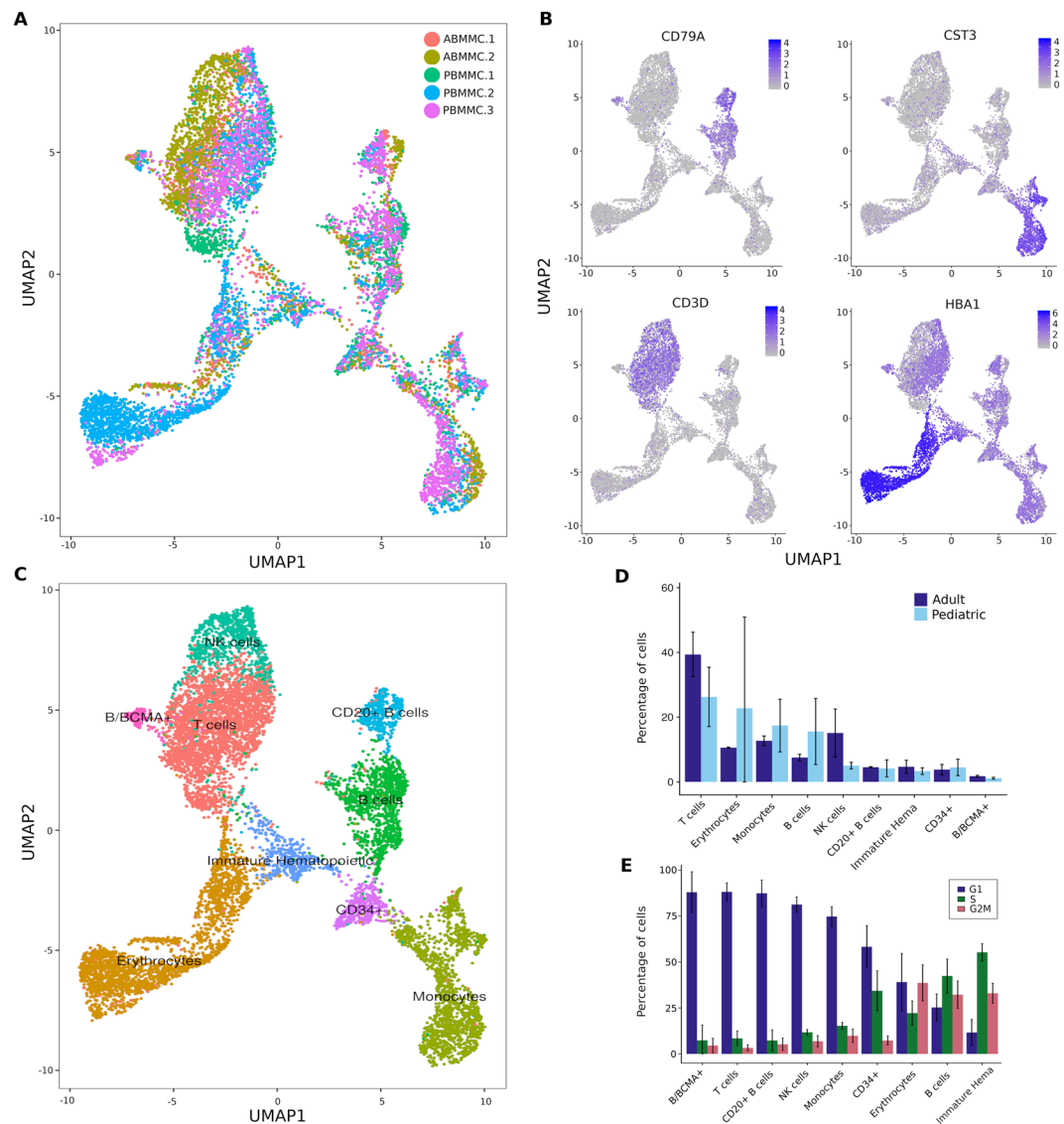
Figure 1. Cell types identified in healthy pediatric and adult bone marrow mononuclear cells (BMMCs) using single cell RNA-seq. (**A**) UMAP representation of healthy BMMCs from three pediatric (PBMMC; n = 6,836 cells) and two adult (ABMMC; n = 3,467 cells) donors. (**B**) Expression of cell surface marker genes used to assign cell types: *CD79A* (B cells), *CST3* (monocytes), *CD3D* (T cells) and *HBA1* (erythrocytes). (**C**) Cell types identified in healthy pediatric and adult BMMCs. (**D**) Proportion of cells of a given cell type in pediatric and adult BMMCs. (**E**) Proportion of healthy cells in predicted cell cycle phases per cell type (G1, S and G2/M).

(Supplementary Figure 2). Subclonal copy number depth ratios were also seen for some of these alterations using exome data (Supplementary Figure 6). We did not observe large cluster specific copy number events for HHD or pre-T samples. We then looked whether there was a correlation between the presence of genetic subclones and intra-individual transcriptional clusters. We used somatic mutations from matched normal and tumor exome data (Supplementary Table 3) and found no major differences in both the number of mutations and genomic locations between samples (except for sample ETV6.RUNX1.1 that had less mutations) (Fig. 4B,C). We identified known cALL mutations in signaling proto-oncogenes (*ABL1, ROS1, NOTCH1*), cell cycle/epigenetic regulators (*CREBBP, CDKN2A*) and noticed few overlapping non-synonymous exonic mutations between samples (Fig. 4D). We used all sufficiently covered (>100×) somatic mutations to generate genetic evolution models and at least one minor genetic subclone was predicted for each sample, with ETV6.RUNX1.2 having two minor subclones (Supplementary Figure 7). We therefore did not observe a correlation between the presence of genetic subclones and observed intra-individual transcriptional heterogeneity, however minor subclones were at the lower end of the allele frequency spectrum and could reflect the neutral evolution frequency tail in some cases[17]. Finally, we looked at somatic mutations at the single cell level using all somatic mutations and retrieved allele calls of individual cells using single cell RNA-seq read alignment files. Around 25% of mutations (n = 668) were covered in at least one cell and ~6.7% (n = 185) of them had at least one cell having the mutated allele (Fig. 4E). A Fisher's exact test was run on mutations having at least >0.1% of cells per sample with the mutated allele (n = 31)

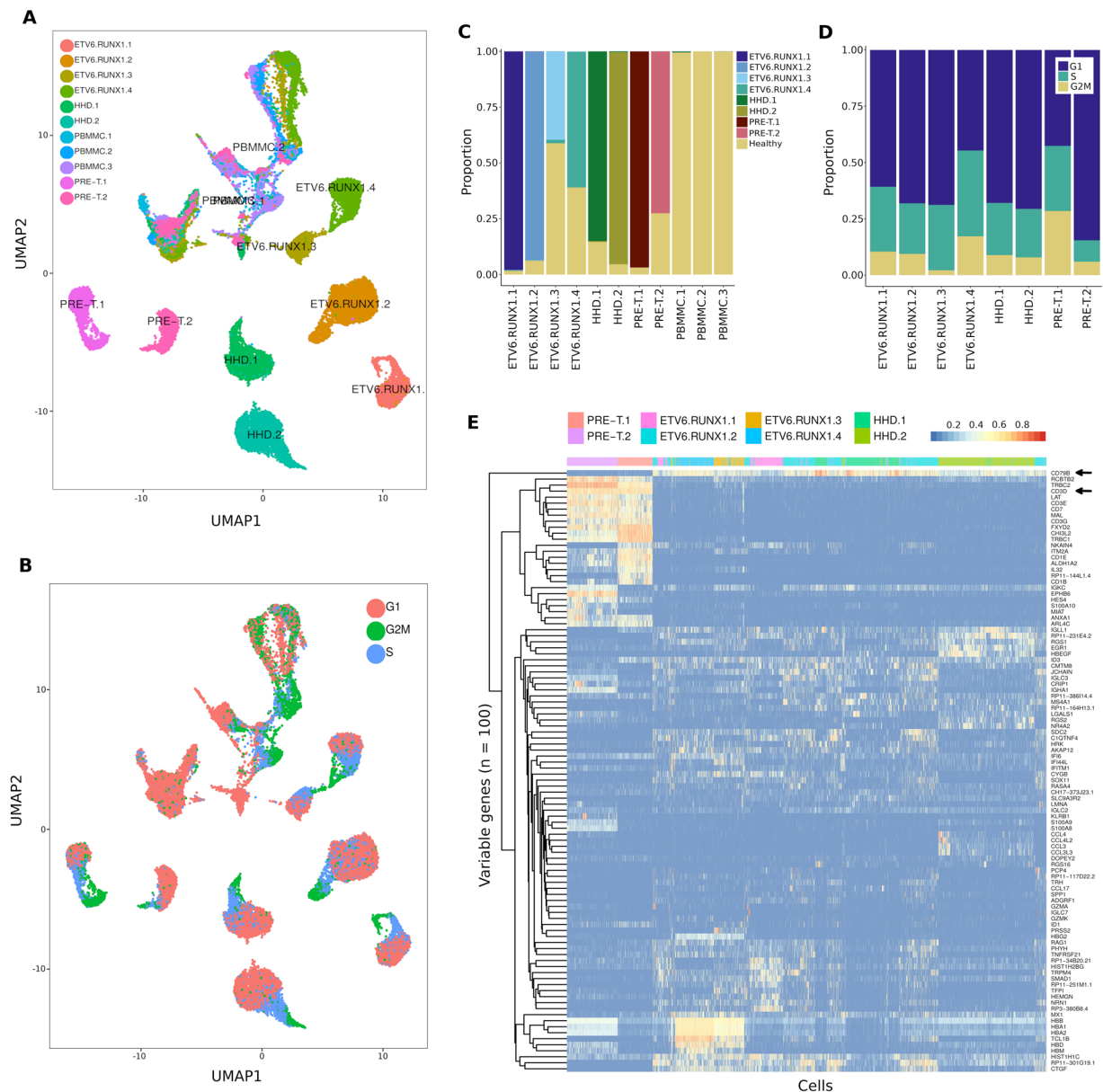**Figure 2.** Transcriptional landscape of cALL cancer cells. (**A**) UMAP representation of BMMCs from three healthy pediatric donors (n = 6,836 cells) and eight cALL patients (n = 32,086 cells). (**B**) UMAP representation of predicted cell cycle phases for healthy and cancer BMMCs. (**C**) Proportion of cells clustering with healthy (PBMMC) cell clusters. (**D**) Proportion of cancer cells in predicted cell cycle phases (G1, S and G2/M). (**E**) Heatmap and unsupervised clustering of normalized and scaled expression of the top 100 most variable genes in leukemia cells.

to identify overrepresented mutations in intra-individual transcriptional clusters. We found somatic mutations in three genes with p-values < 0.1 for samples ETV6.RUNX1.2 and PRE-T.2 (intron of C20orf194, 3′ UTR of SH3 domain containing gene *SH3D21* and exon of pyridoxal kinase gene *PDXK*), indicating a slight enrichment of the mutated allele in a transcriptional cluster (Supplementary Table 4). Overall very few somatic mutations were fulfilling the criteria for statistical testing and many mutations had a low number of observations contributing to underinflation of p-values and non-significance of all mutations after multiple testing correction (Fig. 4F, Supplementary Table 4). Thus, it is still possible that other subclonal somatic mutations could be enriched in specific transcriptional clusters but were not possible to detect using this approach. Overall these results suggest that there is some limited evidence linking genetic alterations detected from bulk exome data to intra-individual transcriptional heterogeneity in these samples.

Developmental states and ribosomal protein gene expression. To address whether the observed intra-individual transcriptional variability could be linked to the maturation and differentiation states of the cancer cells, we implemented two developmental state classifiers based on the expression profiles of healthy B
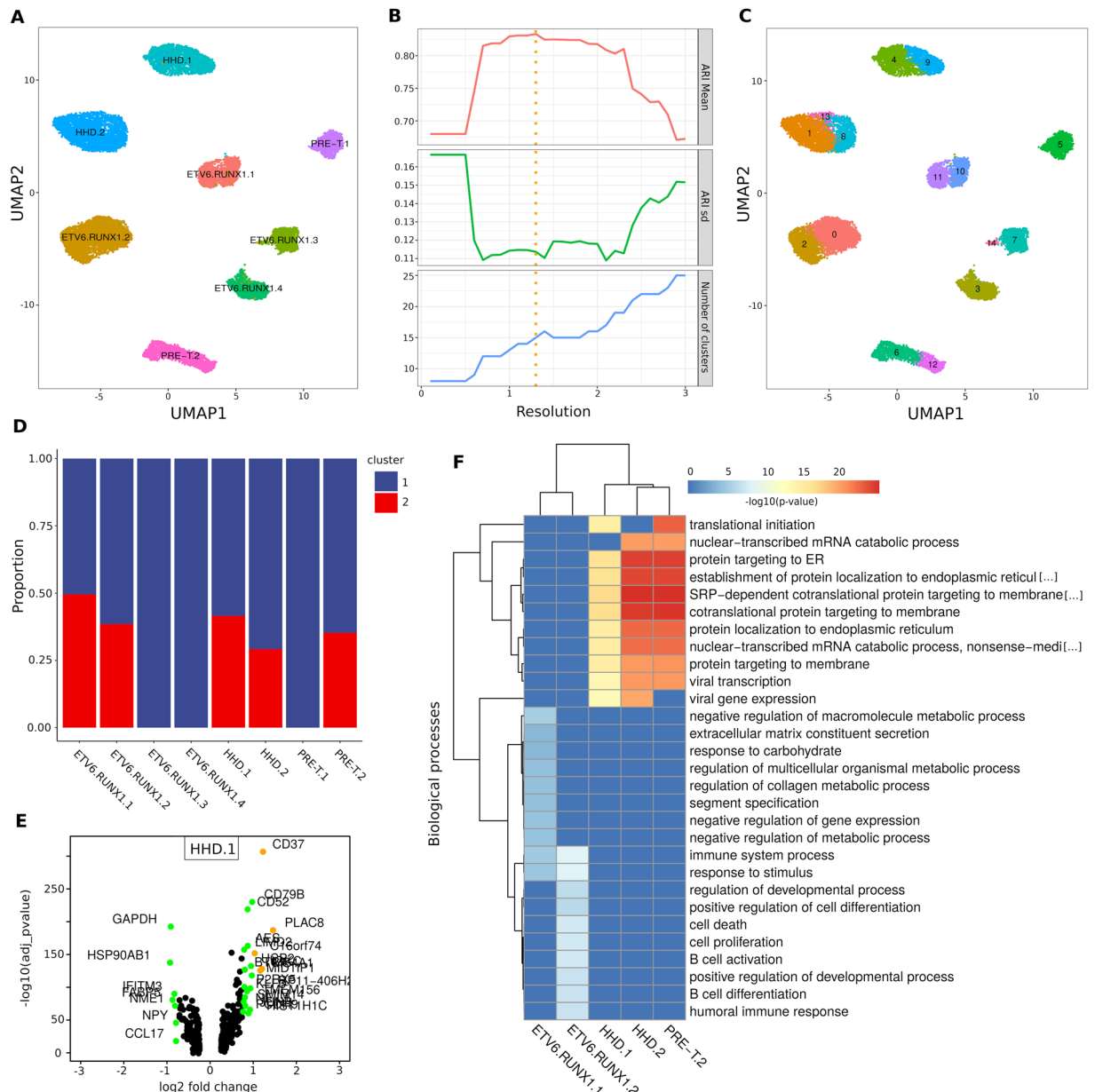
Figure 3. Intra-individual transcriptional heterogeneity reveals deregulated genes and pathways within cALL samples. (**A**) UMAP representation of cALL cells in G1 phase not clustering with healthy cell clusters (n = 16,731). (**B**) Mean Adjusted Rand Index (ARI) of clustering solutions over a range of resolutions (highest mean ARI at 1.3 resolution). (**C**) Clusters of cells identified in cALL samples using the highest mean ARI resolution. (**D**) Proportion of cells belonging to each intra-individual cluster after removing clusters having less than 10% of cells (n = 16,162). (**E**) Differentially expressed genes between two the clusters of cells within the HHD.1 sample (log fold-change > 0.75 = green, >1 = orange). (**F**) Heatmap and unsupervised clustering of enriched GO biological pathways obtained using the top 100 most significant differentially expressed genes of cALL samples.

and T cells. We assigned pseudotime values of 0 to 1 along the maturation spectrum from stem-like to mature developmental states (Fig. 5A).

The performances of the classifiers were assessed on healthy cells using cross-validation and showed good predictive performance using average Root Mean Squared Error (RMSE) (Fig. 5B). Final classifiers were trained on all healthy B and T cells and then used to obtain the predicted developmental states of every cancer cell. We observed statistically significant shifts in developmental pseudotime distributions within samples (Fig. 5C). The most significant shifts were found within HHD and PRE-T samples, while less pronounced shifts were observed within *ETV6-RUNX1* samples. These findings correlate with differential expression results that revealed modulated expression of ribosomal protein genes in HHD and PRE-T samples (Supplementary Table 1). An inverse relationship between maturation states of normal hematopoietic cells and ribosomal protein (RP) expression
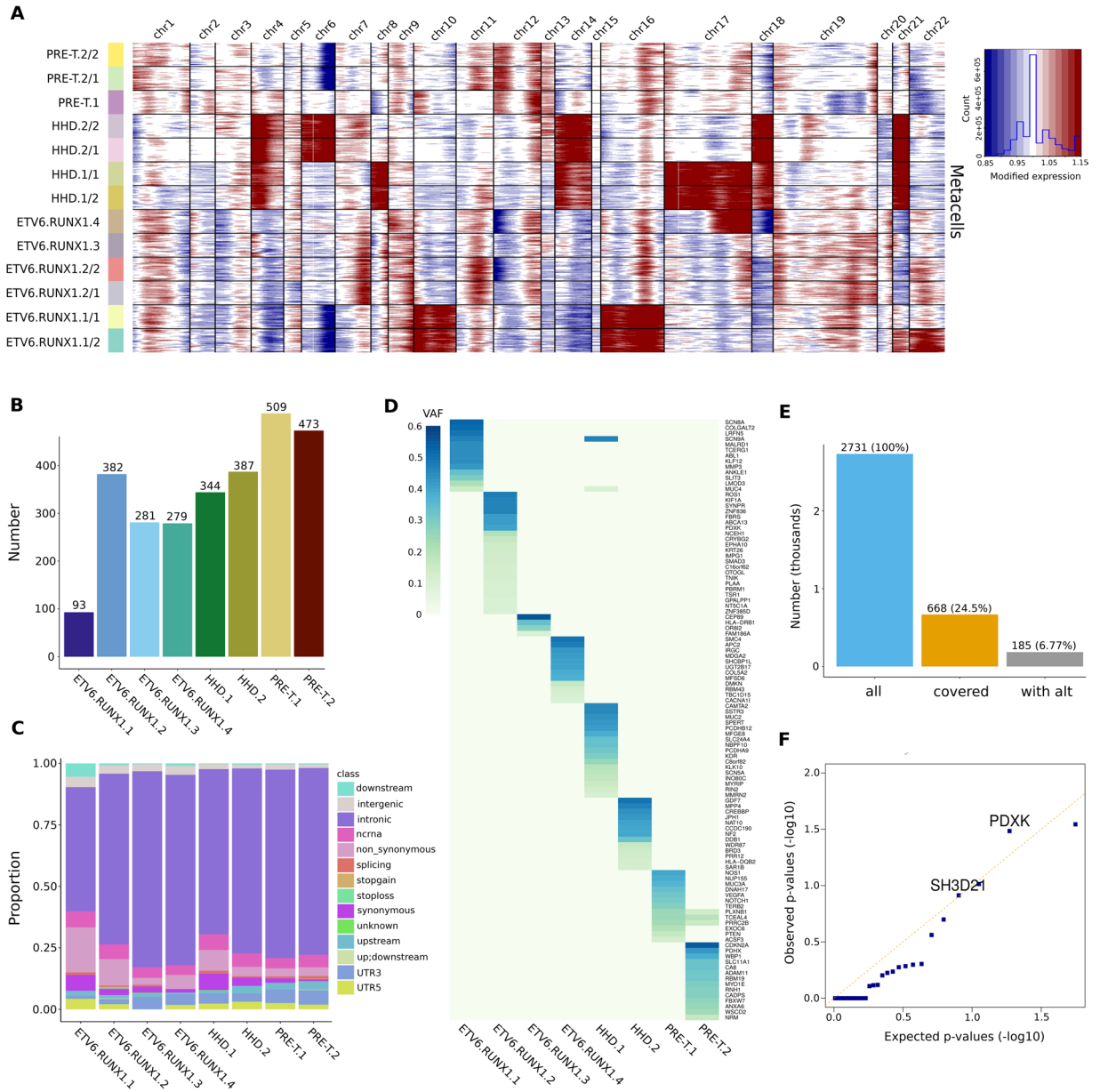
Figure 4. Somatic alterations and intra-individual expression variability. (**A**) Copy number profiles of samples and intra-individual transcriptional clusters using equal size metacells and healthy pediatric BMMCs as control cells. (**B**) Number of somatic mutations in cALL samples. (**C**) Genomic annotations of somatic mutations in cALL samples. (**D**) Gene names and variant allele frequencies of exonic non-synonymous somatic mutations in cALL samples. (**E**) Number of somatic mutations used as input to obtain allele calls from single cell RNA-seq alignment files using vartrix (all = number of GRCh38 mutations lifted from hg19; covered = number of mutations covered in at least one cell; with alt = number of mutations with at least one mutant allele call). (**F**) QQ-plot of Fisher's exact test p-values of somatic mutations tested for enrichment in intra-individual transcriptional clusters (>0.1% of cells with mutant allele call, n = 31 mutations).

levels was previously reported in a zebrafish single cell RNA-seq study[18]. We observed the same relationship for cALL cancer cells in all samples exhibiting intra-individual transcriptional clusters where subsets of cells with higher pseudotime had lower ribosomal protein expression levels (Fig. 5D). This trend was clearer in samples having strong (e.g. HHD.1) vs weak (e.g. ETV6.RUNX1.2) shifts in pseudotime (Fig. 5E). Ordering cells by RP expression revealed a gradient of expression that was correlated to clusters having similar behavior (Fig. 5F). These findings suggest that developmental states and RP expression could contribute to the intra-individual transcriptional heterogeneity observed in some cALL patients.
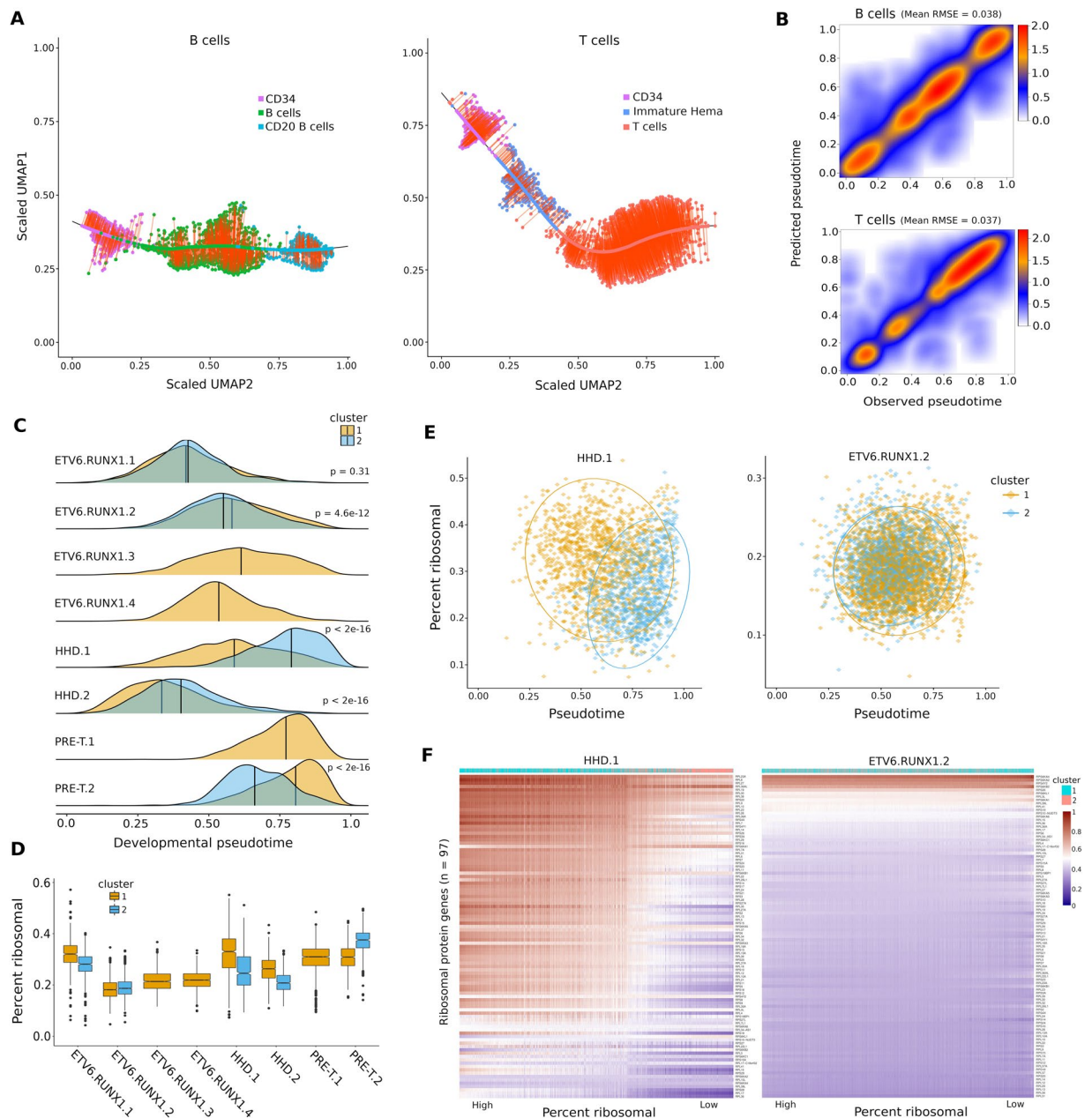
Figure 5. Predicted developmental state is inversely correlated with ribosomal protein expression in cALL cells and is a major source of intra-individual transcriptional heterogeneity. (**A**) Left: UMAP representation of the maturation spectrum of healthy pediatric and adult B cells (CD34+ → B cells → CD20 + B cells) used for the B cell developmental state classifier. Right: UMAP representation of the maturation spectrum of healthy pediatric and adult T cells (CD34+ → immature hematopoietic → T cells) used for the T cell developmental state classifier. Cells were projected onto the loess fit of the spectrum and assigned a pseudotime value of 0 to 1 from the first stem-like cell to the last mature cell. (**B**) Observed vs predicted pseudotime of healthy B and T cells using a hundred 70/30 cross-validation splits; mean RMSE was computed over all splits. (**C**) Density of predicted developmental state pseudotime distributions of leukemia cells per sample and intra-individual transcriptional cluster. (**D**) Boxplot of ribosomal protein (RP) expression percentage in leukemia cells per sample and intra-individual transcriptional cluster. (**E**) Pseudotime vs ribosomal protein expression in samples showing strong (HHD.1) vs weak (ETV6.RUNX1.2) intra-individual pseudotime and ribosomal protein expression shifts. (**F**) Heatmap of normalized and scaled expression of ribosomal protein genes per cell sorted from high to low. An expression gradient correlated to cluster assignment can be observed in sample HHD.1 but not ETV6.RUNX1.2.

Discussion

We applied a single cell expression strategy to uncover major sources of intra-individual transcriptional heterogeneity in childhood ALL. To our knowledge this is the first study on intra-individual transcriptional heterogeneity in cALL patients using single cell gene expression. Our study found two notable patterns at diagnosis: first,

we found transcriptional heterogeneity associated to the predicted developmental state of leukemia cells which was inversely correlated with the expression of ribosomal protein genes. This observation was more obvious in HHD and pre-T samples and less so in *ETV6-RUNX1* samples, indicating possible subtype specificity. The disrupted products of the fusion between the *ETV6* and *RUNX1* transcription factors, involved in B cell maturation, could result in an overall lower maturation potential of the leukemia cells. Second, we observed transcriptional heterogeneity linked to gene expression and metabolic regulation, B cell activation and cell proliferation/death in *ETV6-RUNX1* samples. We found significant joint over expression of the JUN and FOS proto oncogenes in a cluster of cells for sample ETV6.RUNX1.2, suggesting deregulation in stress/cancer genes.

To further understand transcriptional variability within samples, we sought to determine if genetic alterations could explain these intra-individual transcriptional clusters. We called copy number variants (CNVs) in each sample to detect events present specific to transcriptional clusters and found a few large gain and loss events in subsets of cells of *ETV6-RUNX1* samples. Given the high number of genes in these regions, the transcriptional impacts of these copy number alterations remain unknown. We found no correlation between the number of predicted genetic subclones in each sample and the presence of intra-individual transcriptional clusters. All samples were predicted to have two or more genetic subclones but not all samples were found to have intra-individual transcriptional clusters, which could be explained in part by our single cell clustering procedure retaining the 'most representative' solution and not returning all possible transcriptional clusters. Another explanation could be false positive predictions of genetic subclones triggered by the density of mutations found at the lower allele frequency spectrum and arising from neutral tumor evolution. The low number of somatic mutations in pediatric cancers[19], coupled with limited coverage of the exome (versus whole-genome) and lower sequencing depth (~200–300X) compared to targeted sequencing (~1000–1500X) could have contributed to the problem. We then looked for somatic mutation allele enrichment in specific transcriptional clusters by assigning allele calls to each cell using the single cell RNA-seq read alignment files. Since very few mutations had enough coverage and alternate allele calls to be tested for statistical significance, we failed to report strong evidence of mutant allele enrichment. These results could be explained by the limited expression levels of genes in each cell and the uneven coverage over genes inherent to the 3′ single cell library protocol. Assessing DNA and RNA simultaneously in the same cell is a strategy that could reduce the limitations mentioned above; technology to do this exists but was initially limited to a few hundred cells[20]. Newer methods can amplify targeted genomic regions[21] or transcripts of interest[22,23] in thousands of cells to generate both the transcriptional and mutational profiles of individual cells. Additionally, an analytical approach that combines both single cell RNA and DNA generated from the same system, but not the same cells, has been developed to jointly analyze genetic and transcriptional clonality and could be useful to apply in subsequent studies[24]. Taken together we found some evidence, although limited, of subclonal genetic alterations that were linked to intra-individual transcriptional heterogeneity in cALL samples. It is likely that subclonal genetic events altering expression could be more easily identifiable in samples where the main source of transcriptional heterogeneity is not related to the predicted developmental state.

Questions remain about the clinical significance of intra-individual transcriptional heterogeneity in cALL. It is unclear whether subsets of cells at various developmental stages or with variable levels of maturation potential have any fitness advantages during treatment. A previous study exploring the relapse potential of cALL patients at diagnosis projected individual cancer cells to their closest B cell maturation state using mass cytometry and found no correlation between the fraction of cells in a given state and relapse potential[25]. Subsets of cells having deregulated expression of oncogenes (e.g. sample ETV6.RUNX1.2) or genes involved in treatment resistance could be more susceptible to clonal selection. Although no patients in our study have relapsed, future single cell studies using matched diagnosis/relapse should provide valuable information on clonal dynamics. Single cell expression data of matched cells should help identify genes and cell surface markers that gain or lose expression in resistant cells and provide insights to guide clinical interventions.

Methods

Samples.
Study subjects were diagnosed in the Division of Hematology-Oncology at the Sainte-Justine Hospital (Montreal, Canada) and part of the Quebec childhood ALL cohort (QcALL)[26]. Our study cohort consisted of 6 B-ALL and of 2 T-ALL patients (8 males) with a mean age at diagnosis of 5.8 years and a mean blast percentage of 82% (Table 1). All patients were treated under DFCI protocols. Control subjects were patients for whom the diagnosis of ALL was not confirmed after bone marrow puncture (3 males) and with a mean age of 2.7 years (Table 1). Tumoral DNA specimens were collected from bone marrow mononuclear cells (BMMCs) at initial diagnosis and paired normal DNA specimens were collected from peripheral blood or bone marrow samples without blast cells. All samples were isolated using a Ficoll-Paque gradient fragmentation, washed in PBS and DNA was extracted using the Gentra Puregene Blood Kit from QIAGEN according to manufacturer protocol. For single cells, bone marrow mononuclear cells were cryopreserved in FBS with 10% DMSO. The Sainte-Justine Institutional Review Board approved the research protocols and informed consents were obtained from all participating individuals and/or their parents. All experiments have been carried out in accordance with relevant guidelines and regulations.

Single cell RNA sequencing and analyses.
Thawed BMMCs were loaded onto the 10X Genomics Chromium single cell platform (v2 chemistry) at McGill University and Genome Quebec Innovation Center. We aimed for 3,000 cells per sample and targeted 100,000 reads per cell by sequencing each sample on one lane of an Illumina HiSeq 4000 high-throughput sequencer (26v98 b.p. paired-end sequencing) (Table 2).

Sequencing reads were processed using the Cell Ranger v2.1.0 pipeline on Ensembl GRCh38.84[12]. BMMCs from three healthy pediatric controls and two publicly available healthy adult control samples (www.10Xgenomics.com) were analyzed jointly using the canonical correlation analysis (CCA) in Seurat v2.3.2[15]. Genes expressed

in at least 5 cells, cells having at least 200 genes expressed and cells having less than 8% mitochondrial reads were retained for the analysis. Expression estimates of the genes were regressed out for number of unique molecular identifiers (nUMI), percentage of mitochondrial reads and the S and G2/M cell cycle scores. The thousand most variable genes were selected for each sample and pooled to create a unique list of genes being highly variable in at least two samples (n = 883). These genes were kept for the CCA and the top ten canonical correlation vectors were used as input for UMAP v0.2.3[27]. Cell types were identified using the expression of known cell surface marker genes (*CD34*/CD34+, *CD79A*/B cells, *MS4A1*/CD20+ B cells, *CST3*/monocytes, *SPN + KIT*/immature hematopoietic, *HBA1*/erythrocytes, *CD3D*/T cells, *NKG7*/NK cells, *MZB1*/BCMA). A second analysis was done on cells of three healthy pediatric controls and eight cancer samples. The same cell and gene filters as described above were applied and a Seurat object was created by iteratively merging each dataset. A principal component analysis (PCA) was run on the most variable genes (n = 468) and the top twenty principal components were used as input for both UMAP and cluster identification using Louvain's algorithm[28] with a resolution of 0.1. Cells from cancer samples that did not cluster with control cell clusters and those that were in the G1 cell cycle phase were retained for downstream analyses. The most variable genes were identified in these filtered cells (n = 560) and used as input for both PCA and UMAP. The most representative clustering solution was obtained by finding clusters using resolutions of 0.1 to 3.0 with a step of 0.1. For each of these 30 solutions, we computed all pairwise Adjusted Rand Index (ARI) values and retained the clustering solution with the highest ARI mean (optimal resolution = 1.3). Clusters containing less than 10% of cells in a sample were discarded for statistical reasons. Intra-individual differential gene expression analyses were done using the MAST test[29] (Supplementary Table 1). Gene ontology analyses were done using goseq v1.3.0[30] on the top hundred differentially expressed genes ranked by adjusted p-value. Single cell allele calls of somatic mutations derived from exome data were obtained using vartrix v1.1.0[31]. Copy number profiles of intra-individual transcriptional clusters were obtained by generating thirty equal-sized metacells (sum of raw expression counts of single cells) per cluster that were used as input to infercnv.v.0.99.4[32]. Metacells of healthy pediatric BMMCs were used as control.

Exome sequencing and analyses.   Exome sequencing and analyses were performed as previously described[33]; whole exomes were captured using Agilent's SureSelect XT Clinical Research Exome kit per manufacturer's protocol and libraries were sequenced on Illumina high throughput sequencers (HiSeq 2500 or 4000 in paired-end 2 × 75b.p. or 2 × 100b.p.) targeting mean coverage of 300X for tumor and 100X for normal. Raw reads were aligned to the hg19 genome reference using bwa v0.7.7[34], duplicates were marked using Picard v1.107 (http://broadinstitute.github.io/picard/) and indels realigned and bases recalibrated using GATK v3.3.0[35]. Somatic mutations of matched tumor/normal were obtained using mutect v1.1.6[36], filtered for the PASS flag and variant allele frequency >0.05 (n = 2748, Supplementary Table 3). Variants were lifted over using vcf-liftover (https://github.com/liqg/vcf-liftover) for use with single cell RNA-seq data (hg19 → GRCh38, n = 2731/2748, 99.4%). Mutations with high coverage (>100X) were used as input to sciClone v1.1.0[37] and clonEvol v0.99.11[38] to generate clonal evolution models. Copy number calls were obtained using sequenza v2.1.0[39].

Developmental state classifiers and ribosomal protein expression.   UMAP coordinates of healthy BMMCs corresponding to the developmental spectrum of B cells (CD34+, B cells and CD20+ B cells) and T cells (CD34+, immature hematopoietic, T cells) were extracted for input to loess fits. Cells were projected onto the fit line at the closest distance. Outlier cells with distances of more than three standard deviations from the mean were discarded. Cell distances along the fit line were calculated, scaled between 0 and 1 and set as the developmental state pseudotime. Normalized and scaled expression values of the most variable genes determined in the Seurat CCA were extracted and used as input for one-layer fifteen nodes neural networks using the R package nnet v7.3–12. Classifier performances on control cells were assessed using cross-validation of a hundred 70/30 splits and mean RMSE values were computed. Final classifiers were trained on all healthy cells and used to predict the developmental pseudotime of each leukemia cell. Ribosomal protein (RP) expression per cell was defined as the ratio of raw reads mapped to ribosomal protein genes (n = 97, Supplementary Table 5) divided by the total number of raw reads per cell.

Data availability
Single cell RNA-seq data are available in GEO under GSE132509. Code are available on github (https://github.com/sinnettlab/childhood_leukemia_single_cell_expression).

References
1. Mulligan, C. G. Molecular genetics of B-precursor acute lymphoblastic leukemia. *J Clin Invest* **122**, 3407–3415, https://doi.org/10.1172/JCI61203 (2012).
2. Woo, J. S., Alberti, M. O. & Tirado, C. A. Childhood B-acute lymphoblastic leukemia: a genetic update. *Exp Hematol Oncol* **3**, 16, https://doi.org/10.1186/2162-3619-3-16 (2014).
3. Hunger, S. P. & Mulligan, C. G. Acute Lymphoblastic Leukemia in Children. *N Engl J Med* **373**, 1541–1552, https://doi.org/10.1056/NEJMra1400972 (2015).
4. Pui, C. H., Mulligan, C. G., Evans, W. E. & Relling, M. V. Pediatric acute lymphoblastic leukemia: where are we going and how do we get there? *Blood* **120**, 1165–1174, https://doi.org/10.1182/blood-2012-05-378943 (2012).
5. Iacobucci, I. & Mulligan, C. G. Genetic basis of acute lymphoblastic leukemia. *Journal of Clinical Oncology* **35**, 975 (2017).
6. Ma, X. *et al.* Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukemia. *Nature communications* **6**, 6604, https://doi.org/10.1038/ncomms7604 (2015).
7. Ferrando, A. A. & López-Otín, C. Clonal evolution in leukemia. *Nature medicine* **23**, 1135 (2017).

8. Irving, J. *et al.* Ras pathway mutations are prevalent in relapsed childhood acute lymphoblastic leukemia and confer sensitivity to MEK inhibition. *Blood* **124**, 3420–3430, https://doi.org/10.1182/blood-2014-04-531871 (2014).

9. Tzoneva, G. *et al.* Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat Med* **19**, 368–371, https://doi.org/10.1038/nm.3078 (2013).

10. Gawad, C., Koh, W. & Quake, S. R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci USA* **111**, 17947–17952, https://doi.org/10.1073/pnas.1420822111 (2014).

11. Spinella, J. F. *et al.* Mutational dynamics of early and late relapsed childhood ALL: rapid clonal expansion and long-term dormancy. *Blood Adv* **2**, 177–188, https://doi.org/10.1182/bloodadvances.2017011510 (2018).

12. Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049, https://doi.org/10.1038/ncomms14049 (2017).

13. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet*, https://doi.org/10.1038/ng.3818 (2017).

14. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome biology* **18**, 212 (2017).

15. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411 (2018).

16. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20**, 40, https://doi.org/10.1186/s12859-019-2599-6 (2019).

17. Graham, T. A. & Sottoriva, A. Measuring cancer evolution from the genome. *The Journal of pathology* **241**, 183–191 (2017).

18. Athanasiadis, E. I. *et al.* Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nature communications* **8**, 2045 (2017).

19. Grobner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327, https://doi.org/10.1038/nature25480 (2018).

20. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods* **12**, 519–522 (2015).

21. Rodriguez-Meira, A. *et al.* Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. *Molecular cell* **73**, 1292–1305. e1298 (2019).

22. van Galen, P. *et al.* Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281. e1224 (2019).

23. Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019).

24. Campbell, K. R. *et al.* clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome biology* **20**, 54 (2019).

25. Good, Z. *et al.* Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. *Nature medicine* **24**, 474 (2018).

26. Healy, J. *et al.* Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood* **109**, 683–692 (2007).

27. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

28. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).

29. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, 278, https://doi.org/10.1186/s13059-015-0844-5 (2015).

30. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14, https://doi.org/10.1186/gb-2010-11-2-r14 (2010).

31. Petti, A. A. *et al.* Mutation detection in thousands of acute myeloid leukemia cells using single cell RNA-sequencing. *BioRxiv*, 434746 (2018).

32. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401, https://doi.org/10.1126/science.1254257 (2014).

33. Khater, F. *et al.* Molecular Profiling of Hard-to-Treat Childhood and Adolescent Cancers. *JAMA network open* **2**, e192906–e192906 (2019).

34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).

35. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–1303 (2010).

36. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213 (2013).

37. Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS computational biology* **10**, e1003665 (2014).

38. Dang, H. *et al.* ClonEvol: clonal ordering and visualization in cancer sequencing. *Annals of oncology* **28**, 3076–3082 (2017).

39. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* **26**, 64–70 (2014).

Acknowledgements

Author contributions

G.B., D.S and M.C. designed and conceived the study. M.C. and P.S. performed bioinformatic analyses. T.S. and C.R. collected and preserved samples. I.R. and Y.C.W. supervised and operated the 10X Genomics Chromium single cell platform. M.C., G.B., and D.S. wrote and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-64929-x.

Correspondence and requests for materials should be addressed to D.S. or G.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.