Supplementary material:

Supplementary methods

Table S1 Patient characteristics

Table S2 Selected genes - literature

Table S3 Selected genes – differential gene expression

Table S4 Differential gene expression between compartments

Supplementary methods

1 - Single cell RNA sequencing and initial preprocessing

Single-cell mRNA sequencing was performed at Single Cell Discoveries (single-cell sequencing service, the Netherlands) according to an adapted version of the CEL-seq protocol[1] (SORT-seq[2]) with primers described in van den Brink et al.[3] . In short, single cells were FACS sorted into 384-well plates containing primers and Mineral oil (Sigma). After sorting, plates were snap-frozen on dry ice and stored at -80° C. For amplification cells were heat-lysed at 65° C followed by cDNA synthesis using steps based on the CEL-Seq2 protocol[1] performed by robotic liquid handling platforms. After second strand cDNA synthesis, the barcoded material was pooled into libraries of 384 cells and amplified using IVT. Following amplification, library preparation was done according to the CEL-seq2 protocol to generate amplified cDNA libraries, which were barcoded using TruSeq small RNA primers (Illumina). The DNA library was paired-end sequenced on an Illumina Nextseq™ 500, high output, with a 1x75 bp Illumina kit (R1: 26 cycles, index read: 6 cycles, R2: 60 cycles)

2 - Initial preprocessing and quality control

The raw reads were preprocessed as follows: Illumina sequencing adapters were removed and reads containing homopolymers were trimmed using cutadapt (version 4.7)[4] . Ribosomal reads were discarded using ribodetector (version 0.3.1)[5] . The preprocessed reads were aligned to the human reference genome (GRCh38, Ensembl release 98) using STARsolo (version 2.7.11b)[6] , and count tables were generated. Reads mapping to multiple locations in the genome were discarded.

The downstream processing of the single cell RNA sequencing data was performed in R (version 4.5.0)[7] using the Seurat (version 5.5.0)[8] . The dataset was filtered to remove any cells with fewer than 1200 total UMIs, more than 6000 unique features, or more than 30% of mitochondrial RNA content. Lastly, the raw counts were normalized using Seurats SCTransform while regressing out any residual signal from mitochondrial transcript percentage.

3- Annotation check of cells and removal based on clusters

To validate the annotation carried out on the flow cytometry data, unsupervised clustering was done to compare the labels from flow cytometry to the transcriptomic clusters. The SCT transformed dataset was subjected to principal component analysis (PCA) on the 3000 most variable genes identified by SCTransform.

To validate the flow cytometry-based annotation, Louvain clustering was performed on the 20 first principal components (PCs) at a resolution of 0.8. Clusters were annotated based on their predominant cell type into either blasts, T-cells or B-cells. All cells with conflicting labels between the flow cytometry and the single cell transcriptomic analyses were discarded. This filtering step removed 31 flow-annotated blasts, 12 healthy B-cells and 29 healthy T-cells. For all subsequent downstream analyses the healthy control cells were disregarded.

4 – Cell cycle scoring

Cells were classified as S or G2/M phase using Seurat's CellCycleScoring on log-normalized counts, using the gene set from Tirosh et al[9], with cells scoring low for both modules classified as G1. The proliferation fraction for each compartment in each patient was calculated as the proportion of cells in S or G2/M phase (Figure 2E).

5 - Dimensionality reduction and assessment of transcriptional separation

To visualize the variance in the data across the patient, compartment and phase parameters, the 30 first PCs were used for a Uniform Manifold Approximation and Projection (UMAP). UMAP parameters were set to a minimum distance of 0.7 and a number of neighbors of 5.

To evaluate transcriptional separation of compartments, a partial least squares discriminant analysis (PLS-DA) was performed on the 30 first PCs using the MixOmics framework (version 6.34.0)[10] . The response variable was set to the compartment label, and the PLS-DA model was fitted with 2 latent components. Decision boundaries were estimated using a centroid distance classifier, which assigns observations to the class of the nearest centroid. Centroid distances were plotted using PLS-DA on the first 30 PCs from the three compartments.

An additional measure for the compartment separation was established using silhouette scores. Euclidean distances were calculated between all cells in PCA space using the first 30 PCs. Silhouette widths were computed using the cluster package (version 2.1.8.2)[11] with compartment labels as the cluster assignment, and these were visualized for each patient separately stratified by compartment.

6 – Differential expression analysis and gene set enrichment analysis

The raw count gene expression matrices for each compartment and patient combination were pseudobulk aggregated. Differential expression analyses compared the CSF samples with the other compartment samples. Thus, a design matrix was constructed based on the patient variable as well as a binary compartment indicator (CSF vs. non-CSF). Differential expression analysis was then carried out using DESeq2 (version 1.50.2)[12] . Contrasts were established comparing the CSF compartment with the other compartments in a paired setting for each patient. Lastly, p-values were corrected for multiple testing using Benjamini-Hochberg. Subsequent gene set enrichment analysis (GSEA),using the fgsea[13] was carried out with FDR correction.

Additionally, differentially expressed genes between primary and relapse leukemia were identified solely in the CSF compartment using DESeq2. Expression was pseudobulked on a patient basis resulting in seven pseudobulked samples, of which 3 were from primary leukemia and 4 were from relapse leukemia. Contrasts were established as the leukemia type and p-values were once again adjusted for multiple testing using the Benjamini-Hochberg method.

7 - Data integration and clustering

The transcriptomics dataset was integrated using CCA integration from the Seurat package using the first 40 PCs. This CCA embedding was used for a Louvain clustering with a resolution of 0.5, as well as the creation of a UMAP.

8 – Trajectory inference and gene expression along pseudotime

Trajectory inference was then carried out using Slingshot (version 2.18.0)[14] with the CCA embeddings as the input and the cluster with the highest proportion of CSF cells as the starting point. This identified 5 different trajectories that were projected from the CCA space onto the UMAP for visualization purposes. The trajectory spanning all three compartments with no dependence on proliferating cells was selected for downstream analysis (Figure 3B). The remaining trajectories were excluded from further analyses because they were based on the cell cycle phase, or they did not encompass all three compartments of interest.

To identify differentially expressed genes across the trajectory, a generalized additive model (GAM) was fitted to the 3000 most variable genes using the tradeSeq package (version 1.24.0)[15] , and a Wald test was implemented in accordance with the recommended slingshot pipeline. P-values were adjusted for multiple testing using Benjamini-Hochberg[16] and adjusted p-values < 0.05 were considered significant.

The 80 significant genes with the highest Wald statistic were selected for visualization. The corresponding expression matrix was corrected for patient effects using the Limma package (version 3.66.0)[17] . The pseudotime parameter was split into 60 equally spaced bins, and the corrected expression matrix was aggregated along these bins before being scaled as Z-scores and visualized alongside the percentwise compositions of patient, compartment, and cell cyclephase classifications for each bin (Figure 3D).

9 – Gene set overrepresentation analysis.

For all of the significant genes, a spearman correlation was calculated for the SCTransformed expression along the pseudotime axis. The genes were then split into two groups: positive ($cor > 0.1$) and negative ($cor < -0.1$) based on the correlation coefficient. These separate gene lists were then utilized for a gene set overrepresentation analysis performed by the ClusterProfiler package (version 4.18.4)[18] using the Gene Ontology, Biological Processes database[19] . The background was set to the 3000 HVGs used for the GAM.

Table S1

Clinical features and cell counts of patients, for whom single cell sequencing of matched cerebrospinal fluid (CSF), bone marrow (BM) and peripheral blood (PB) was performed. For two patients, there were no leukemic cells in peripheral blood.

Patient	**Age**	**Subtype**	**Sample**	**Cytogenetic subtype**	**Sequenced cells, BM**	**Sequenced cells, PB**	**Sequenced cells, CSF**
1	15	B-ALL	Diagnosis	BCR::ABL1	676	304	153
2	4	B-ALL	Diagnosis	ETV6::RUNX1	603	307	10
3	3	T-ALL	Diagnosis	No identified aberrations	304	315	220
4	7	B-ALL	Relapse	B-other, poor risk CNA	601	318	266
5	15*	B-ALL	Diagnosis/ Relapse	Hypodiploid	271	No leukemic cells in PB	235
6	17	B-ALL	Relapse	iAMP21	90	No leukemic cells in PB	596
7	6	T-ALL	Relapse	No identified aberrations	343	314	48

* For patient 5, CSF was sampled at time of isolated CNS relapse, and bone marrow was sampled at time of diagnosis, where there was no CNS involvement. Age is from time of diagnosis.

Table S2

Genes with experimental evidence for a role in CNS invasion and/or CNS+ relapse.

Gene	Evidence	Reference
CD99	Involved in T-ALL adhesion to meningeal cells and chemoresistance	20,21
IL7R	Anti-IL7R monoclonal antibody treatment strongly reduced leukemic infiltration of the CNS in a BCP-ALL xenograft model, IL7R expression in BM BCP-ALL was found to be significantly elevated in CNS+ compared with CNS− cases BCP-ALL cells retrieved from the CSF of children with isolated CNS relapse showed a significantly higher IL7R expression compared with ALL cells from BM at diagnosis and BM at BM relapse without CNS involvement	22
VEGF-A	VEGF-A expression is upregulated in ALL cells in the CSF as a result of adaptation to hypoxia.. VEGF-A is important for transendothelial migration and leptomeningeal infiltration by ALL cells as well as disruption of BBB integrity during leukemic invasion of the brain parenchyma. CSF level of VEGF-A is a good predictor of CNS involvement	23–26
CXCR4	CXCR4/CXCL12 chemokine axis contributes to leukemia-meningeal cell adhesion. Inhibition of CXCR4 prevents CNS involvement otherwise observed using Jurkat model cell line in mice. Studies of human T-ALL demonstrated that CNS relapsed T-ALL occurs in part through the CXCR4-CXCL12 chemokine pathway	27–29
CCR7	CCR7 blockade with a monoclonal antibody almost abrogated homing of acute lymphoblastic leukemia cells to the central nervous system in xenografts, and in a Notch1-driven T-ALL mouse model CCR7 expression was necessary and sufficient for CNS infiltration. In patients, CCR7/CXCR4 expression correlated with central nervous system involvement in T-cell acute lymphoblastic leukemia.	30,31
ITGA6	Mice with ALL xenografts were treated with either a PI3Kδ inhibitor, which decreased α6 integrin expression on ALL cells, or specific α6 integrin-neutralizing antibodies and showed significant reductions in ALL transit along bridging vessels, blast counts in the cerebrospinal fluid and CNS disease symptoms	32
ITGA4	integrin-deficient T-ALL accumulated in the CNS compared to control T-ALL. Mechanistically, loss of integrins accelerated T-ALL proliferation in the CNS, suggesting that integrin-mediated interactions may promote quiescence in this space	33
CD79a	CD79a (also known as Igα) associated with CNS-infiltration in BCP-ALL patients and downregulation of CD79a hampers the engraftment of leukemia cells in different murine xenograft models, particularly in the CNS.	34
JUN	Significantly higher levels of the AP-1 gene, jun proto-oncogene, in patients initially diagnosed as CNS-positive BCP-ALL compared with CNS-negative cases as well as CNS-relapse vs non-CNS-relapse cases in a cohort of 100 patients with BCP-ALL.	35

ZAP70	ZAP70 knockdown in ALL cell lines reduced CCR7/CXCR4 expression via ERK and impaired CNS infiltration in NSG mice, while ZAP70 upregulation increased CCR7/CXCR4 expression and migration towards CCL19 and CXCL12 gradients. In 130 B-cell precursor acute lymphoblastic leukemia and 117 T-cell acute lymphoblastic leukemia patients, zeta-chain-associated protein kinase 70 and CCR7/CXCR4 expression levels were significantly correlated	[31]
SPP1	In vitro functional evidence supports a role for SSP1 in CNS entry by showing that a splice variant of SSP1 favours matrix invasion, adhesion and chemoresistance SPP1 is overexpressed in B-ALL cells present in BM at primary diagnosis of patients who will develop a CNS relapse.	[36,37]
SCD1	a metabolic signature of fatty-acid synthesis in CNS leukemia, with Stearoyl-CoA desaturase (SCD1) as a key player. In vivo SCD1 overexpression increases CNS disease, whilst genetic or pharmacological inhibition of SCD1 decreases CNS load	[38]
RAC2, LGMN, ITGAL, ICAM1	RAC2, AEP /LGMN, ITGAL and ICAM1 expression are associated with CNS disease in a mouse model of pre-B childhood acute lymphoblastic leukemia	[39]

Table S3

Differential expression of selected genes (n=16, Table S2) in cerebrospinal fluid (CSF) versus matched, combined bone marrow (BM) and peripheral blood (PB) after pseudobulked aggregation on sample level (suppl. Methods, section 6).

BaseMean: mean of normalized counts across all samples; Log2FoldChange: $\log_2$ fold change in CSF relative to BM/PB (positive values indicate higher expression in CSF); LfcSE: standard error of the $\log_2$ fold change; Stat: Wald test statistic (Log2FoldChange/LfcSE); Pvalue: Two-sided Wald test P value. P values are not corrected for multiple testing, given the restricted, pre-specified gene set (Table S2).

Gene	BaseMean	Log2FoldChange	LfcSE	Stat	Pvalue
ITGA6	99,72235	-0,37965	0,372456	-1,0193	0,308059
ITGA4	665,5891	-0,09839	0,194366	-0,50624	0,612692
ITGAL	41,11275	-0,16742	0,500732	-0,33435	0,738114
VEGFA	14,10419	1,795481	0,701687	2,558805	0,010503
ZAP70	60,20689	-0,04023	0,178184	-0,22577	0,821381
SPP1	6,549517	3,309629	1,201627	2,75429	0,005882
SCD	44,47332	2,481677	0,563158	4,406712	1,05E-05
JUN	87,24587	-2,29403	0,510897	-4,4902	7,12E-06
CXCR4	703,6821	0,831978	0,281734	2,953059	0,003146
CCR7	27,78254	0,649591	0,873354	0,743789	0,457004
CD79A	713,0683	-0,41509	0,34417	-1,20606	0,227793
RAC2	362,9068	0,019098	0,191796	0,099576	0,920681
LGMN	29,31921	0,454959	0,780108	0,5832	0,559759
ICAM1	4,016464	0,658218	0,655574	1,004033	0,315362
CD99	1153,778	-0,4824	0,214771	-2,2461	0,024698
IL7R	248,6689	0,447375	0,374248	1,195399	0,231931

Table S4 (separate suppl. document in .xls format)

Full list of differentially expressed genes between cerebrospinal fluid (CSF) and matched, combined bone marrow (BM) and peripheral blood (PB) after pseudobulked aggregation on sample level (suppl. Methods, section 6). BaseMean: mean of normalized counts across all samples; Log2FoldChange: $\log_2$ fold change in CSF relative to BM/PB (positive values indicate higher expression in CSF); LfcSE: standard error of the $\log_2$ fold change; Stat: Wald test statistic (Log2FoldChange/LfcSE); Pvalue: Two-sided Wald test P value. Padj: FDR corrected p-value.

1. Hashimshony, T. *et al.* CEL-Seq2 : sensitive highly-multiplexed. *Genome Biol.* 1–7 (2016) doi:10.1186/s13059-016-0938-8.

2. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385-394 e3 (2016).

3. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).

4. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10 (2011).

5. Deng, Z. L., Münch, P. C., Mreches, R. & McHardy, A. C. Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.* **50**, E60 (2022).

6. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *Preprint* Preprint at https://doi.org/10.1101/2021.05.05.442755 (2021).

7. R Core Team. *R: A Language and Environment for Statistical Computing. Available from: Https://Www.R-Project.Org/*. (R Foundation for Statistical Computing, 2025).

8. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).

9. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (1979).* **352**, 189–196 (2016).

10. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, (2017).

11. Maechler, M. , R. P. , S. A. , H. M. , H. K. *Cluster: Cluster Analysis Basics and Extensions. R Package Version 2.1.8.2.* (2026).

12. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).

13. Korotkevich, G. *et al.* Fast gene set enrichment analysis. Preprint at https://doi.org/10.1101/060012 (2016).

14. Street, K. *et al.* Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, (2018).

15. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).

16. Yoav Benjamini, Y. H. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* **57**, 289–300 (1995).

17. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

18. Xu, S. *et al.* Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**, 3292–3320 (2024).

19. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* vol. 25 25–29 Preprint at https://doi.org/10.1038/75556 (2000).

20. Jonart, L. M. *et al.* Disrupting the leukemia niche in the central nervous system attenuates leukemia chemoresistance. *Haematologica* **105**, 2130–2140 (2020).

21. Ebadi, M., Jonart, L. M., Ostergaard, J. & Gordon, P. M. CD99 antibody disrupts T-cell acute lymphoblastic leukemia adhesion to meningeal cells and attenuates chemoresistance. *Sci. Rep.* **11**, 24374 (2021).

22. Alsadeq, A. *et al.* IL7R is associated with CNS infiltration and relapse in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood* **132**, 1614–1617 (2018).

23. Kato, I. *et al. Letters to Blood Hypoxic Adaptation of Leukemic Cells Infiltrating the CNS Affords a Therapeutic Strategy Targeting VEGFA*. http://ashpublications.org/blood/article-pdf/129/23/3126/1400733/blood721712.pdf.

24. Münch, V. *et al.* Central nervous system involvement in acute lymphoblastic leukemia is mediated by vascular endothelial growth factor. *Blood* **130**, 643–654 (2017).

25. Kinjyo, I., Bragin, D., Grattan, R., Winter, S. S. & Wilson, B. S. Leukemia-derived exosomes and cytokines pave the way for entry into the brain. *J. Leukoc. Biol.* **105**, 741–753 (2019).

26. Tang, Y.-T., Jiang, F., Guo, L., Si, M.-Y. & Jiao, X.-Y. Expression and significance of vascular endothelial growth factor A and C in leukemia central nervous system metastasis. *Leuk. Res.* **37**, 359–66 (2013).

27. Jonart, L. M. *et al.* CXCR4 antagonists disrupt leukaemia-meningeal cell adhesion and attenuate chemoresistance. *Br. J. Haematol.* **201**, 459–469 (2023).

28. Walker, K. L. *et al.* CXCR4 allows T cell acute lymphoblastic leukemia to escape from JAK1/2 and BCL2 inhibition through CNS infiltration. *Leuk. Lymphoma* **62**, 1167–1177 (2021).

29. Jost, T. R. *et al.* Role of CXCR4-mediated bone marrow colonization in CNS infiltration by T cell acute lymphoblastic leukemia. *J. Leukoc. Biol.* **99**, 1077–1087 (2016).

30. Buonamici, S. *et al.* CCR7 signalling as an essential regulator of CNS infiltration in T-cell leukaemia. *Nature* **459**, 1000–1004 (2009).

31. Alsadeq, A. *et al.* The role of ZAP70 kinase in acute lymphoblastic leukemia infiltration into the central nervous system. *Haematologica* **102**, 346–355 (2017).

32. Yao, H. *et al.* Leukaemia hijacks a neural mechanism to invade the central nervous system. *Nature* https://doi.org/10.1038/s41586-018-0342-5 (2018) doi:10.1038/s41586-018-0342-5.

33. Lux, S. Y. *et al.* Integrin-deficient T cell leukemia accumulates in the central nervous system. *bioRxiv* https://doi.org/10.1101/2024.06.28.601274 (2026) doi:10.1101/2024.06.28.601274.

34. Lenk, L. *et al.* CD79a promotes CNS-infiltration and leukemia engraftment in pediatric B-cell precursor acute lymphoblastic leukemia. *Commun. Biol.* **4**, 73 (2021).

35. Gebing, P. *et al.* A brain organoid/ALL coculture model reveals the AP-1 pathway as critically associated with CNS involvement of BCP-ALL. *Blood Adv.* **8**, 4997–5011 (2024).

36. Santoro, J. C., Bastos, A. C. S. F., Gimba, E. R. P. & Emerenciano, M. Reinforcing osteopontin as a marker of central nervous system relapse in paediatric B-cell acute lymphoblastic leukaemia: SPP1 splice variant 3 in the spotlight. *Br. J. Haematol.* **186**, e88–e91 (2019).

37. van der Velden, V. H. J. *et al.* New cellular markers at diagnosis are associated with isolated central nervous system relapse in paediatric B-cell precursor acute lymphoblastic leukaemia. *Br. J. Haematol.* **172**, 769–781 (2016).

38. Savino, A. M. *et al.* Metabolic adaptation of acute lymphoblastic leukemia to the central nervous system microenvironment depends on stearoyl-CoA desaturase. *Nat. Cancer* **1**, 998–1009 (2020).

39. Holland, M. *et al.* RAC2, AEP, and ICAM1 expression are associated with CNS disease in a mouse model of pre-B childhood acute lymphoblastic leukemia. *Blood* **118**, 638–49 (2011).