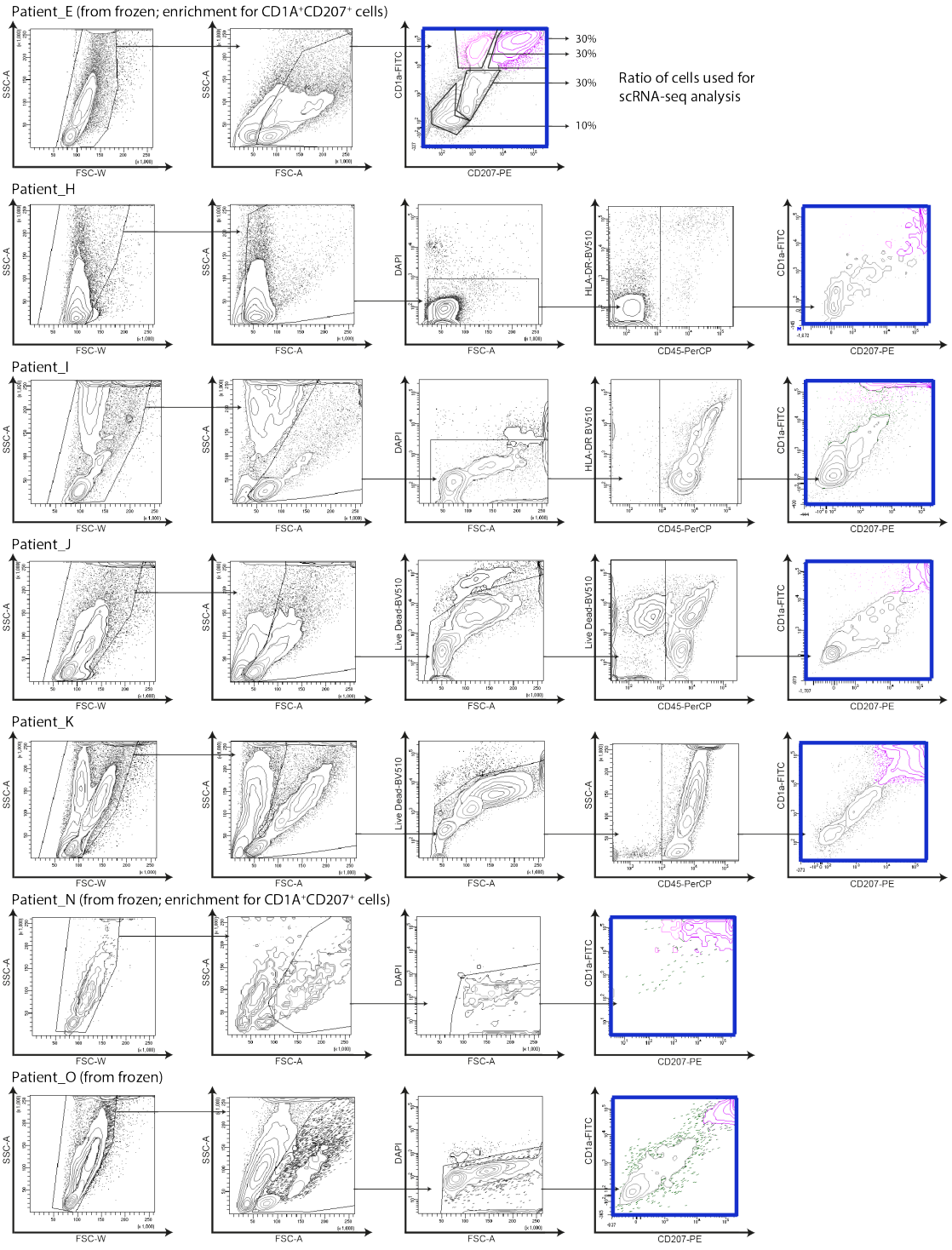
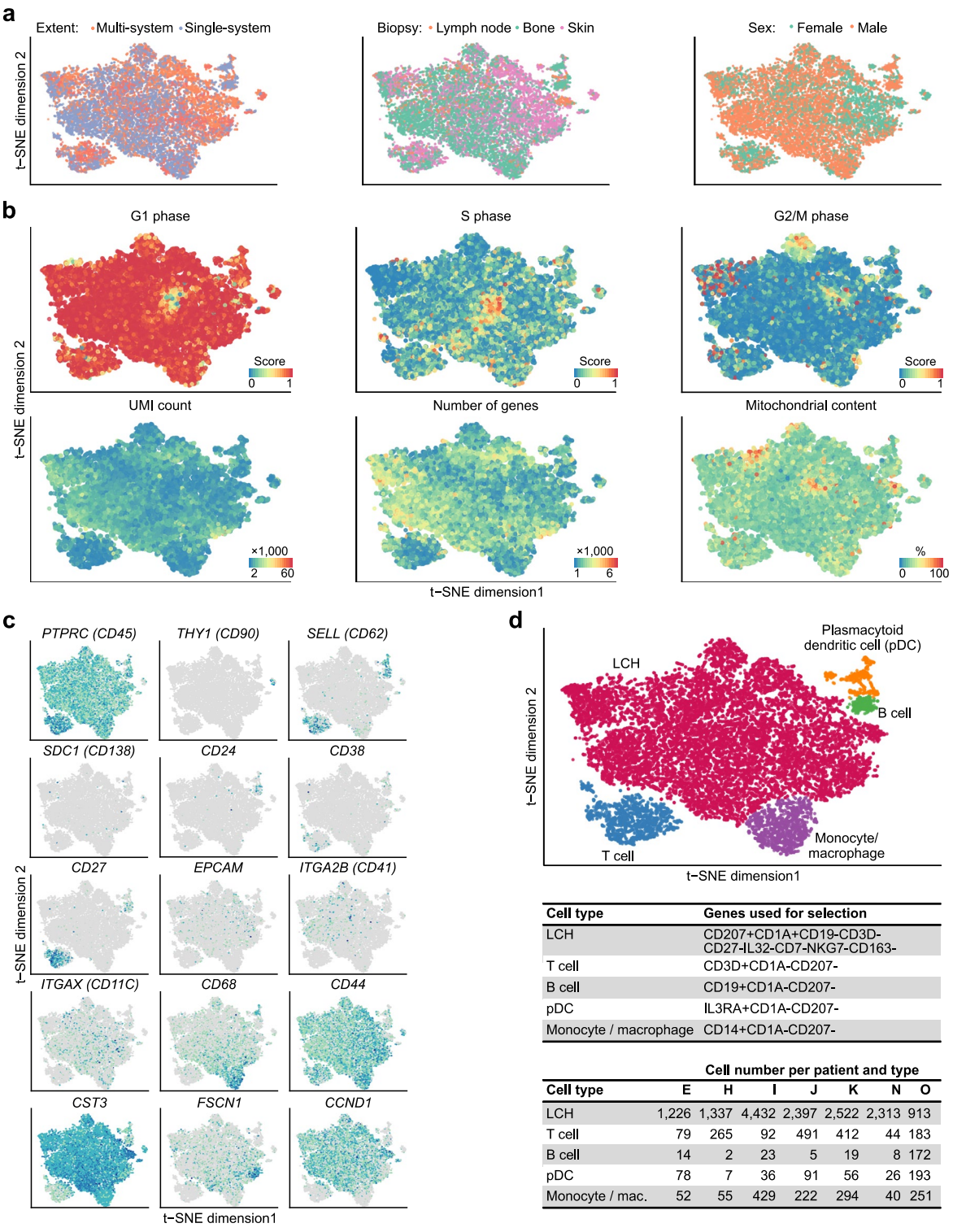


Supplementary Figure 1: Cellular heterogeneity in LCH lesions based on flow cytometry. Flow cytometry plots of LCH biopsy samples used for single-cell transcriptome profiling. Blue frames show sorted cells used as input for the single-cell RNA-seq experiments. Each panel corresponds to a different patient sample. In patient LCH_E, four cell populations with varying degrees of CD1A and CD207 expression were used in different percentages for single-cell RNA-seq (as indicated); in patient LCH_N the gates for forward scatter and side scatter were set such that the majority of lymphocytes were excluded.



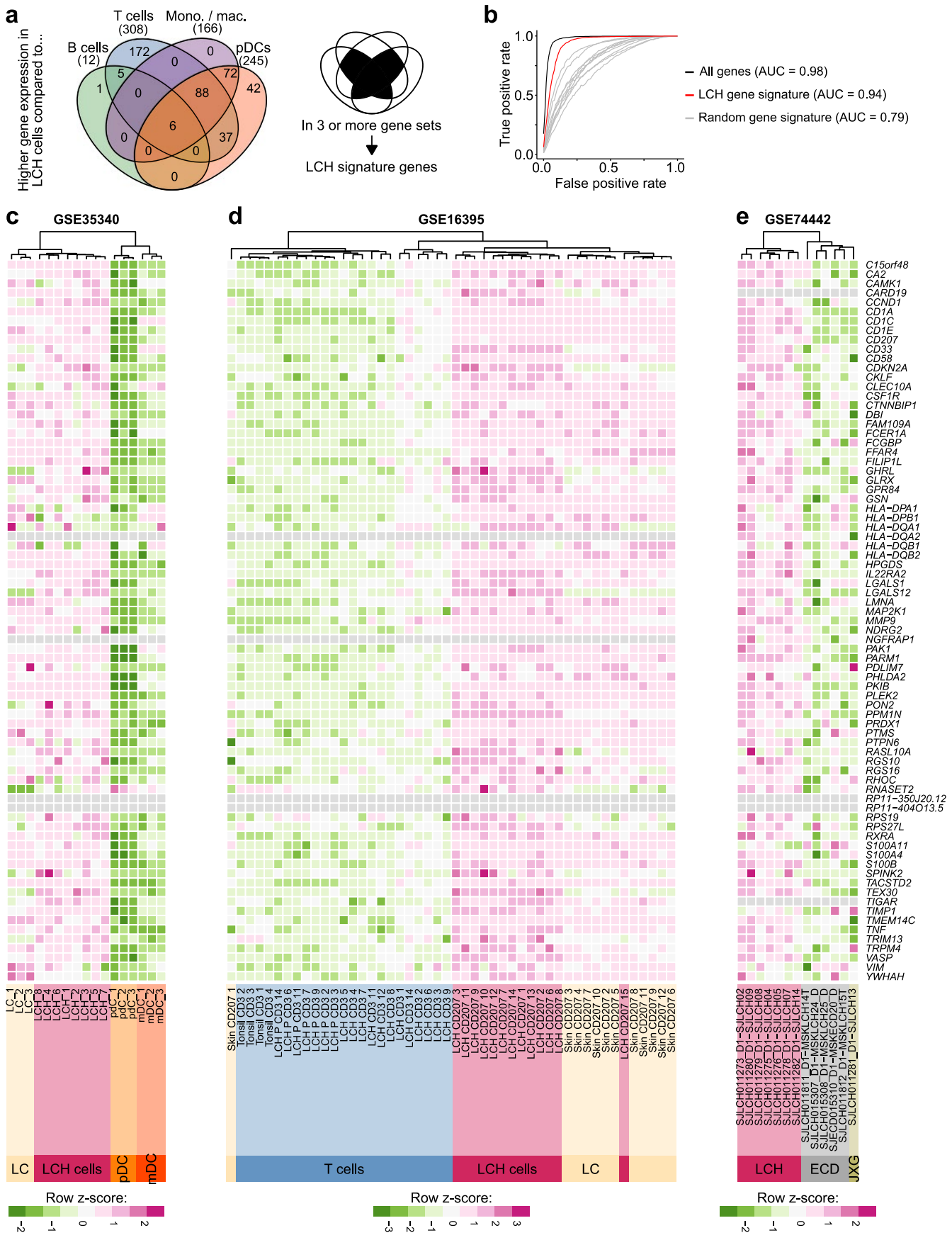
Cell type	Genes used for selection
LCH	CD207+CD1A+CD19-CD3D-CD27-IL32-CD7-NKG7-CD163-
T cell	CD3D+CD1A-CD207-
B cell	CD19+CD1A-CD207-
pDC	IL3RA+CD1A-CD207-
Monocyte / macrophage	CD14+CD1A-CD207-

Cell number per patient and type

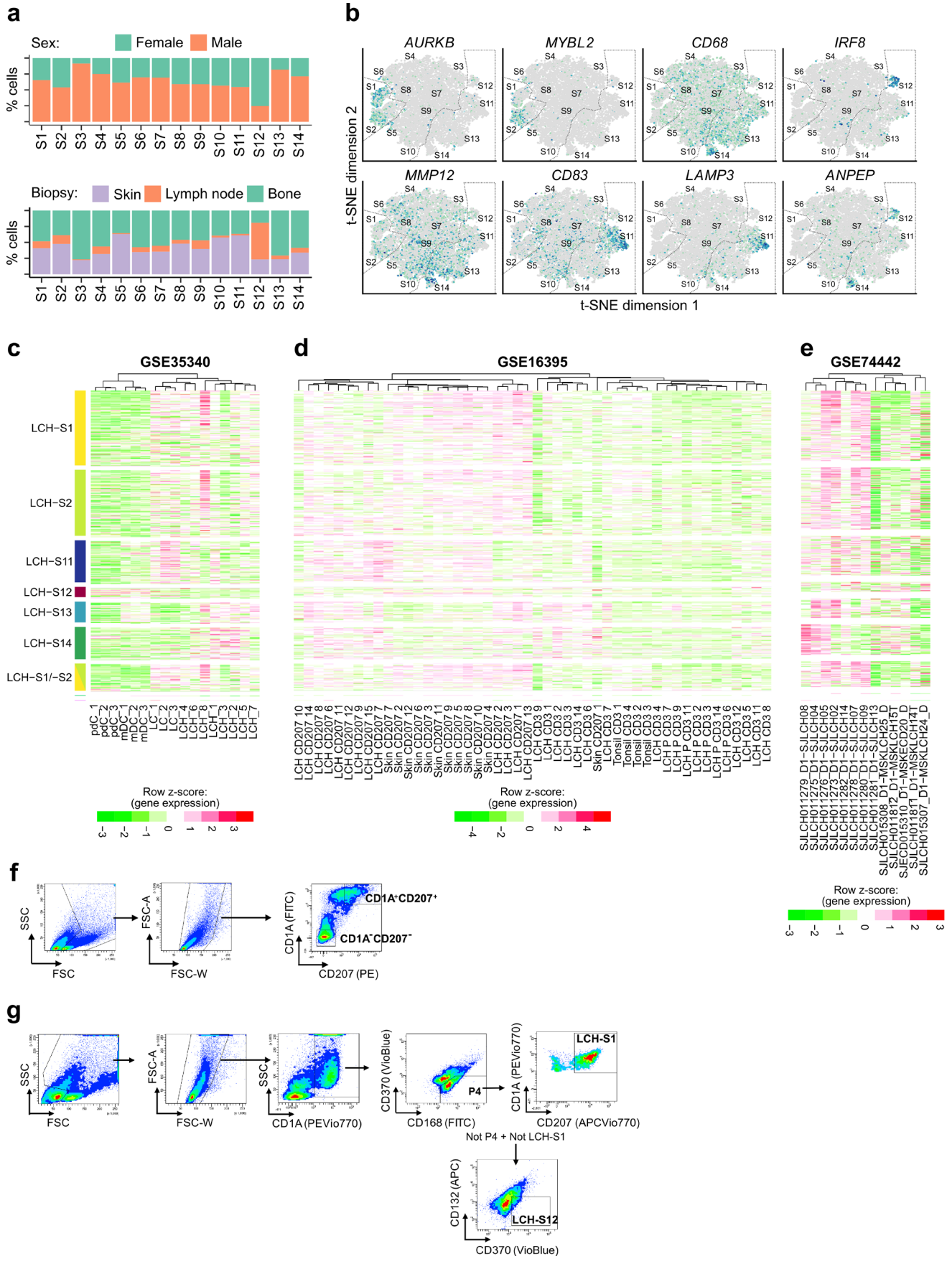
Cell type	E	H	I	J	K	N	O
LCH	1,226	1,337	4,432	2,397	2,522	2,313	913
T cell	79	265	92	491	412	44	183
B cell	14	2	23	5	19	8	172
pDC	78	7	36	91	56	26	193
Monocyte / mac.	52	55	429	222	294	40	251

Supplementary Figure 2: Cell type labelling and data characteristics. (a-c) t-SNE plots overlaid with information about the corresponding biopsy sample (panel **a**) as well as different data metrics (panel **b**; high values are shown in red, low values in blue), and with gene expression levels for specific immune cell types (panel **c**; colours indicate relative UMI counts, with darker colours corresponding to stronger expression). **(d)** t-SNE plot (top) overlaid with labels for LCH cells and four immune cell populations assigned based on the expression of canonical marker genes (middle). Total cell numbers per patient and cell type are shown in a table (bottom).

Supplementary Fig. 3

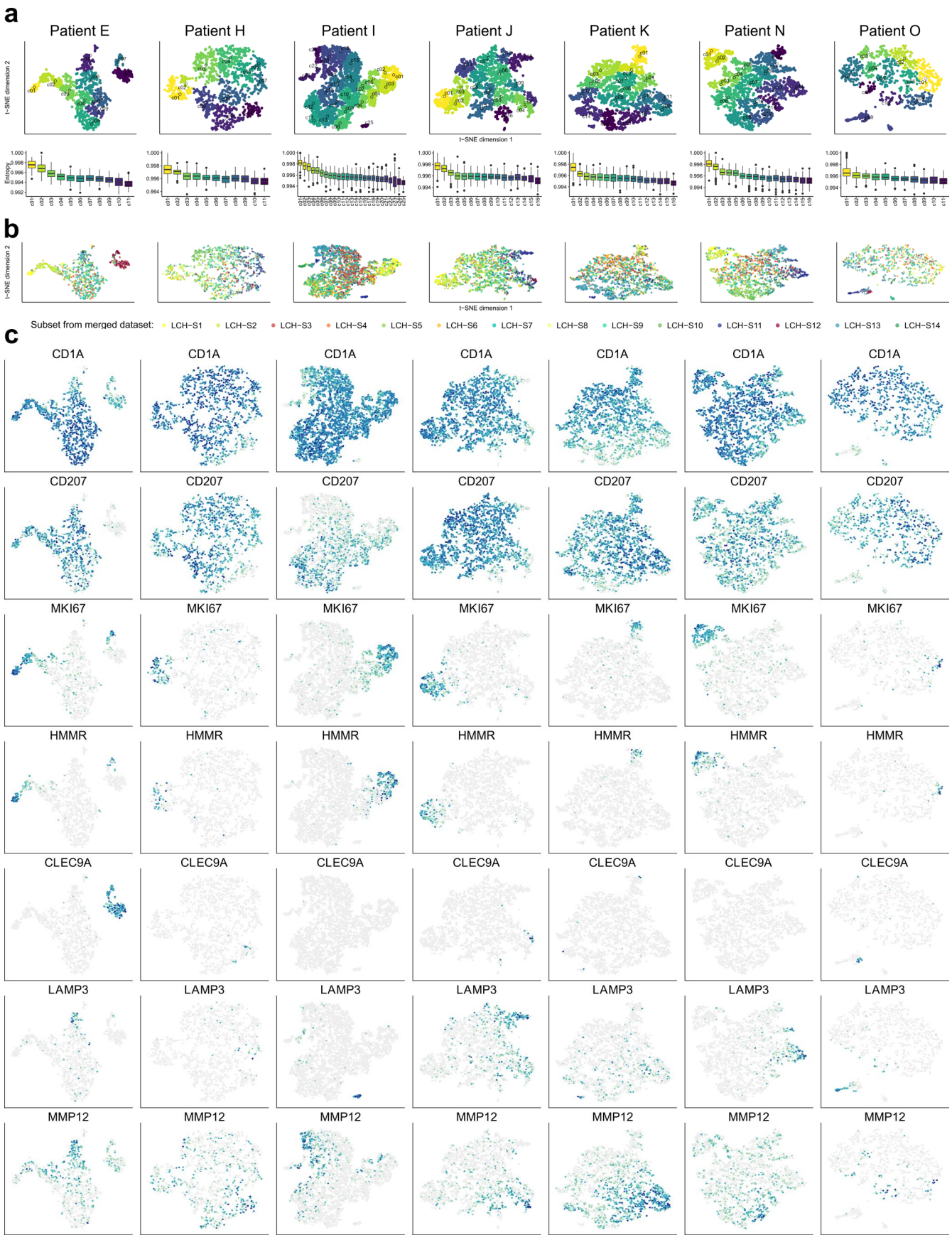


Supplementary Figure 3: Gene expression signature of LCH cells applied to external datasets. (a) Venn diagram showing the overlap of genes upregulated in LCH cells in comparison to (from left to right) B cells, T cells, monocytes / macrophages, and plasmacytoid dendritic cells (pDCs) in our dataset. Genes founds in at least three out of four comparisons were included in the gene signature of LCH cells. **(b)** Receiver operating characteristic (ROC) curve showing the performance of a generalized linear model classifier for the discrimination of LCH cells from non-LCH cells. We used the *glmnet* package in R ("binomial" family, default parameters) and trained the models on the 19,044 single-cell transcriptomes in our dataset. Three types of models were compared differing by the genes used for training: all measured genes (black), only genes in the LCH signature (red), or genes in ten sets of 77 random genes (same number of genes as in the LCH gene signature). The mean performance in 10-fold cross-validation is shown and the area under the ROC curve (AUC) is indicated. **(c)** Heatmap of gene expression values for the LCH gene signature in external microarray data obtained from GEO (accession: GSE35340). Colours indicate relative gene expression per row. LC, healthy epidermal Langerhans cell; pDC, healthy plasmacytoid dendritic cells; mDC, healthy myeloid dendritic cells 1 (i.e., $CD1c^+$ DCs). **(d)** Same as panel **c** with expression data from accession GSE16395. Skin CD207, healthy epidermal Langerhans cells; Tonsil CD3, $CD3^+$ cells healthy isolated from tonsil; LCH P CD3, $CD3^+$ cells isolated from the blood of patients with LCH; LCH CD3, $CD3^+$ cells isolated from LCH lesions; LCH CD207, $CD207^+$ cells isolated from LCH lesions. **(e)** Same as panel **c** with expression data from accession GSE74442. Samples are from seven patients with LCH, five patients with Erdheim-Chester disease (ECD), and one patient with juvenile xanthogranuloma (JXG).

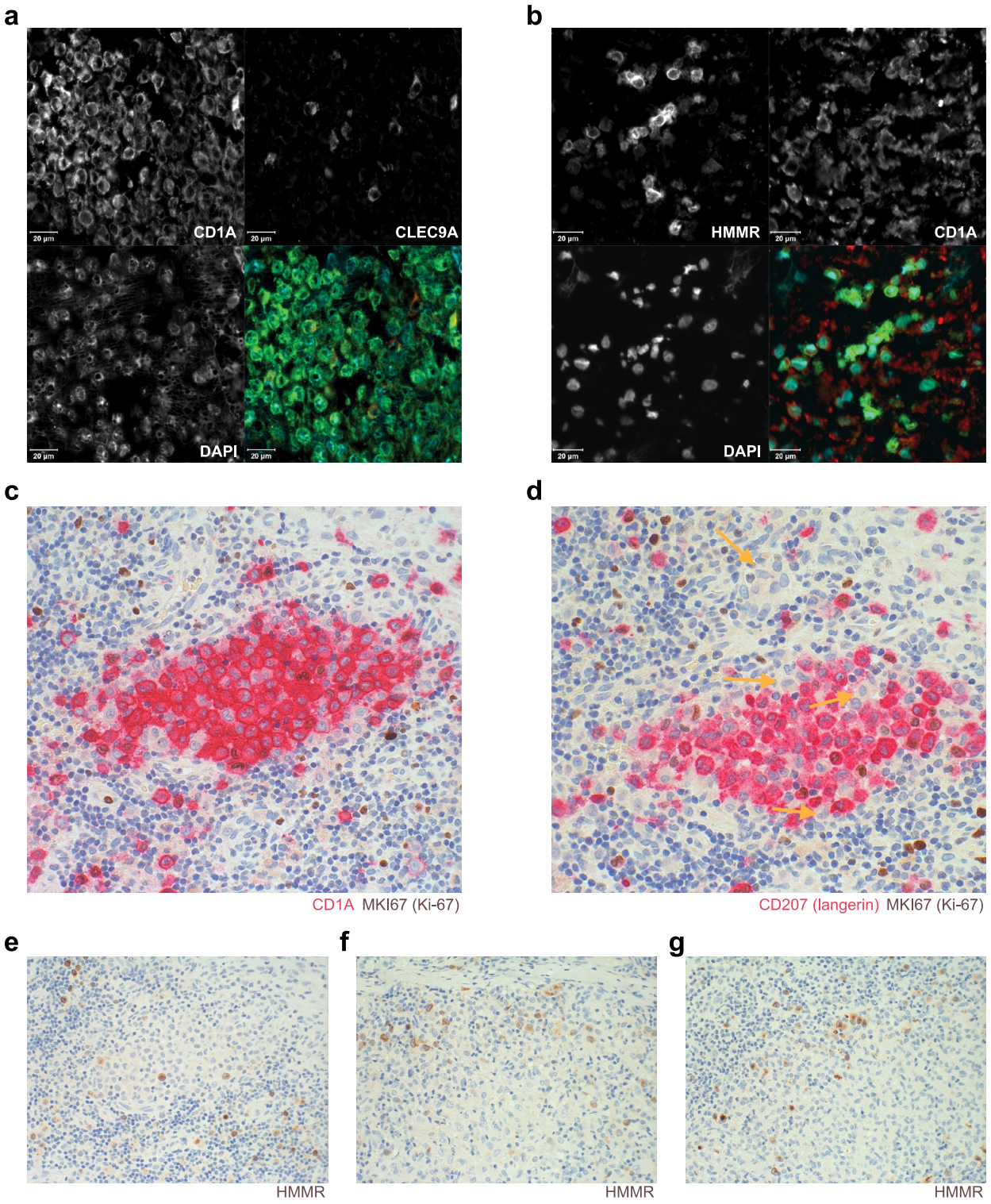

a
Sex:
Female
Male
% cells
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
Biopsy:
Skin
Lymph node
Bone
b
AURKB
MYBL2
CD68
IRF8
MMP12
CD83
LAMP3
ANPEP
t-SNE dimension 2
t-SNE dimension 1
c
GSE35340
LCH−S1
LCH−S2
LCH−S11
LCH−S12
LCH−S13
LCH−S14
LCH−S1/−S2
Row z-score:
(gene expression)
d
GSE16395
Row z-score:
(gene expression)
e
GSE74442
Row z-score:
(gene expression)
f
SSC
FSC
FSC-A
FSC-W
CD1A⁺CD207⁺
CD1A⁻CD207⁻
CD1A (FITC)
CD207 (PE)
g
SSC
FSC
FSC-A
FSC-W
SSC
CD1A (PEVio770)
CD370 (VioBlue)
CD168 (FITC)
P4
CD1A (PEVio770)
LCH-S1
CD207 (APCVio770)
Not P4 + Not LCH-S1
CD132 (APC)
LCH-S12
CD370 (VioBlue)

Supplementary Figure 4: LCH subset markers in external data and FACS enrichment of subsets. (a) Bar plots indicating the relative distribution of male and female cells (top), and cells obtained from biopsies in different anatomical locations (bottom) in each of the 14 LCH cell subsets. **(b)** t-SNE plots showing expression levels of selected marker genes. Colours indicate relative UMI counts with darker colours corresponding to stronger expression. **(c)** Heatmap of gene expression values for LCH subset gene signatures in external microarray data obtained from GEO (accession: GSE35340). Colours indicate relative gene expression per row. LC, healthy epidermal Langerhans cell; pDC, healthy plasmacytoid dendritic cells; mDC, healthy myeloid dendritic cells 1 (i.e., $CD1c^+$ DCs). **(d)** Same as panel **c** with expression data from accession GSE16395. Skin CD207, healthy epidermal Langerhans cells; Tonsil CD3, healthy $CD3^+$ cells isolated from tonsil; LCH P CD3, $CD3^+$ cells isolated from the blood of patients with LCH; LCH CD3, $CD3^+$ cells isolated from LCH lesions; LCH CD207, $CD207^+$ cells isolated from LCH lesions. **(e)** Same as panel **c** with expression data from accession GSE74442. Samples are from seven patients with LCH, five patients with Erdheim-Chester disease (ECD), and one patient with juvenile xanthogranuloma (JXG). **(f)** Flow cytometry plots showing the gating strategy for sorting of non-LCH cells ($CD1A^-CD207^-$) and bulk LCH cells ($CD1A^+CD207^+$) from patient LCH_E. **(g)** Flow cytometry density plots showing the gating strategy to enrich LCH-S1 and LCH-S12 LCH cell subsets using CD132 as an additional negative marker for LCH-S12 and exclusion of $CD207^-$ cells from LCH-S1 from patient LCH_E.

Supplementary Fig. 5

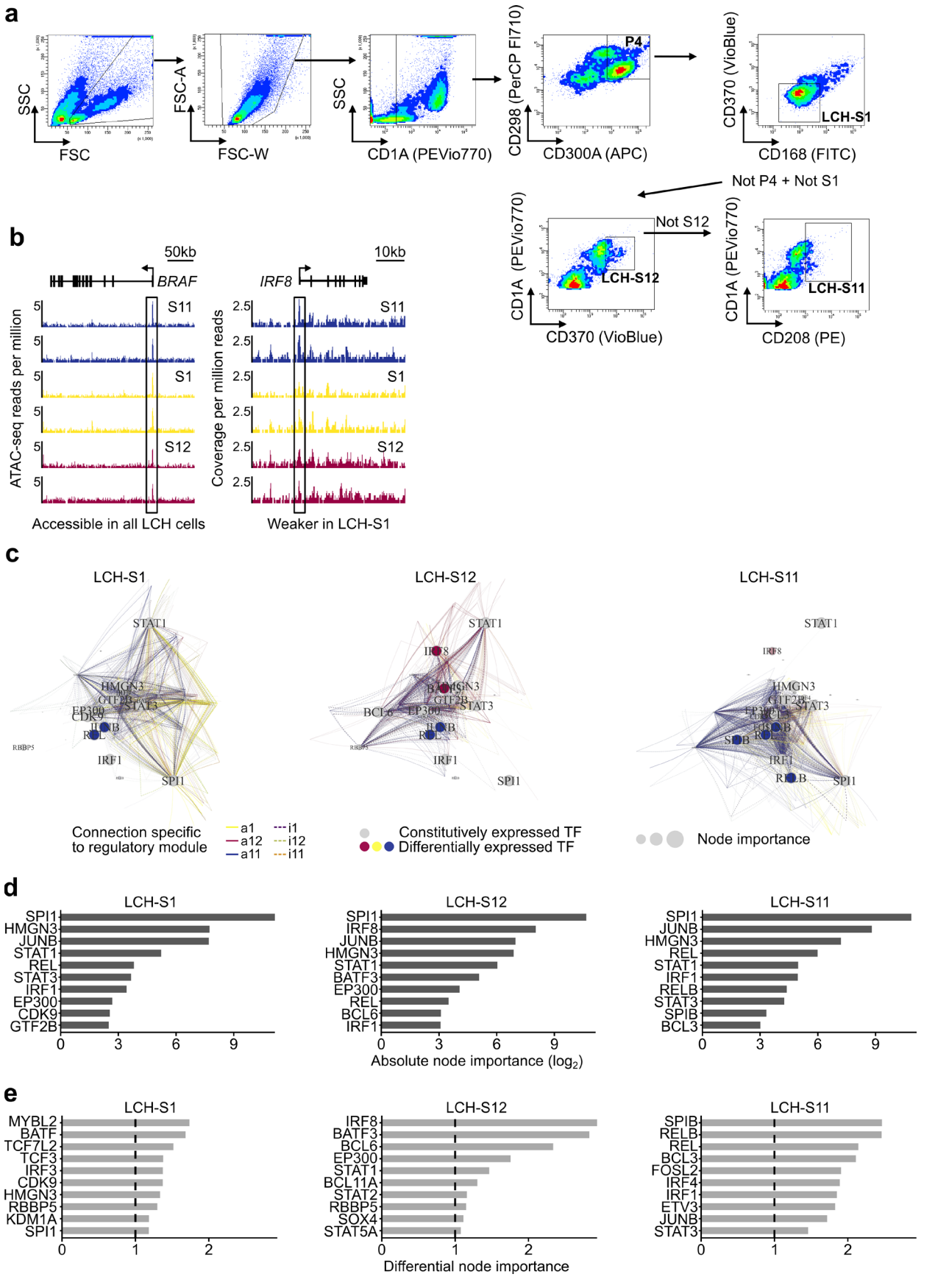


Supplementary Figure 5: Single-cell entropy and gene expression for individual single-cell RNA-seq datasets. (a) t-SNE plots (top) based on single-cell transcriptomes from seven LCH patient biopsies individually (from left to right). Cells were clustered algorithmically and coloured by descending signalling entropy, as shown in the bar plots below. **(b)** t-SNE plots as in panel **a**, overlaid with annotation about the LCH subset each cell was assigned to in the combined dataset (**Figure 3a**). **(c)** t-SNE plots showing expression levels of selected marker genes. Colours indicate relative UMI counts with darker colours corresponding to stronger expression.



Supplementary Figure 6: Immunohistochemistry and immunofluorescence of LCH cell subsets. (a,b) Representative immunofluorescence images of LCH lesions stained for CD1A and CLEC9A (panel **a**), HMMR and CD1A (panel **b**), and DAPI (both). **(c)** LCH lesion stained for CD1A (red) and Ki-67 (brown). Nuclei were counterstained with haematoxylin. **(d)** LCH lesion (same lesion as in panel **c** stained for CD207 (red) and Ki-67 (brown). Yellow arrows point to putative LCH cells (identified by morphology) that do not stain for CD207. **(e-g)** Immunohistochemistry staining for HMMR of three LCH lesions. Nuclei were counterstained with haematoxylin.

Supplementary Fig. 7



Supplementary Figure 7: LCH cell subset signatures and characteristics in five LCH samples. (a) Flow cytometry plots showing the gating strategy to enrich LCH cell subsets from patient LCH_E. LCH-S1: positive for HMMR (CD169); LCH-S12: positive for CLEC9A (CD370); LCH-S11: positive for LAMP3 (CD208). These samples were used for the ATAC-sequencing analysis. **(b)** Genome browser plots showing normalised ATAC-seq signal intensity at two gene loci. The BRAF promoter is accessible in all LCH cell subsets examined (left), whereas IRF8 shows reduced accessibility in LCH-S1 cells (right). **(c)** Gene regulatory networks inferred for three LCH cell subsets (LCH-S1, LCH-S12, LCH-S11) based on single-cell transcriptome and ATAC-seq data, and the key regulators identified in **Figure 5d,e**. Nodes in the network correspond to enriched transcription factors as well as their putative target genes (based on sequence proximity and chromatin 3D structure). Node size is proportional to both gene expression level and node out-degree (i.e., number of outgoing connections from the transcription factor) in the respective LCH subset. In this network, the gene expression values from seven LCH patients have been combined (in contrast to **Figure 6a** which only used data from patient LCH_E). Edge colours indicate the module of the corresponding peak, and edge visibility is proportional to chromatin accessibility. The network layout was automatically generated using the *igraph* package. **(d)** Bar plots showing the top-10 transcription factors ranked by node importance (calculated as a combination of gene expression level and node out-degree) in the networks in panel **c**. **(e)** Bar plots showing the top-10 transcription factors ranked by differential node importance (node importance in one network relative to the mean across all three networks) in the networks in panel **c**.

Supplementary Material

Supplementary Table 1. Data overview. List of sequenced libraries for single-cell RNA-seq and bulk ATAC-seq with summary statistics.

Supplementary Table 2. Marker genes and genomic regions. Lists of differentially expressed genes between LCH cells and non-LCH immune cells, marker genes for each LCH cell subset, and differentially accessible genomic regions between LCH cell subsets.

Supplementary Table 3. Enrichment analysis. Detailed results of the enrichment analyses of genes and genomic region sets.

Supplementary File 1. Gene regulatory networks. Interactive diagrams for the web-browser-based exploration of LCH-subset-specific gene regulatory networks.

Supplementary Website: http://LCH-hierarchy.computational-epigenetics.org.