Supplementary Figures

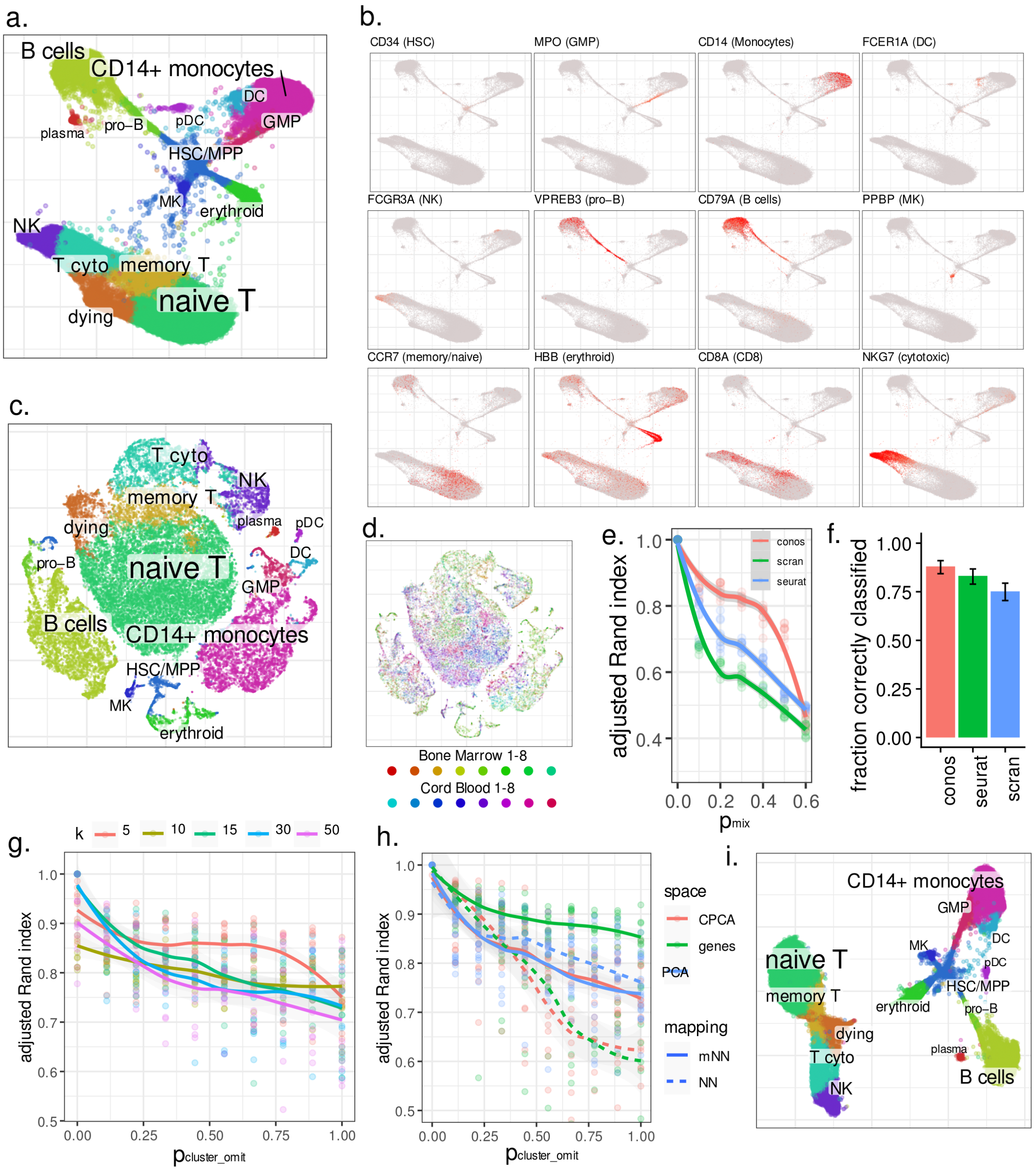
Joint analysis of heterogeneous single-cell RNA-seq dataset collections

Authors: Nikolas Barkas[1*], Viktor Petukhov[1,2*], Daria Nikolaeva[1], Yaroslav Lozinsky[1], Samuel Demharter[2], Konstantin Khodosevich[2], and Peter V. Kharchenko[1,3,†]

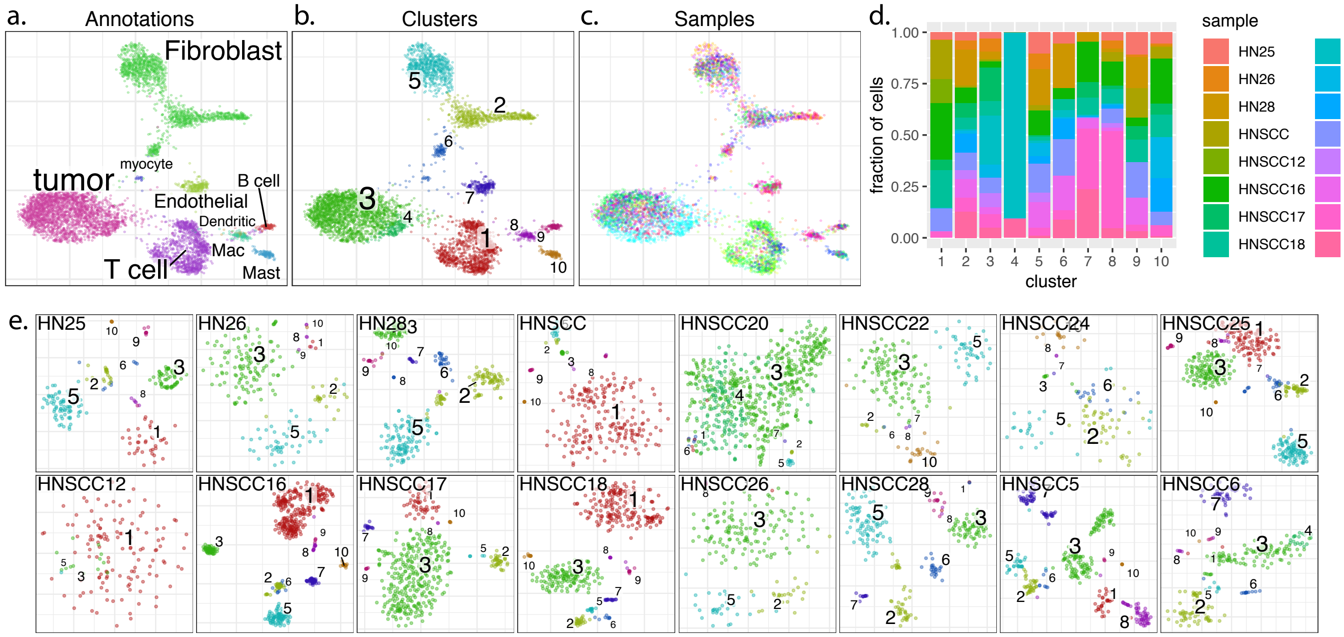
Affiliations: [1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115 USA. [2]Biotech Research and Innovation Centre (BRIC), Faculty of Health, University of Copenhagen, Copenhagen N, DK-2200 Denmark. [3]Harvard Stem Cell Institute, Cambridge, MA 02315 USA.

* The authors have contributed equally

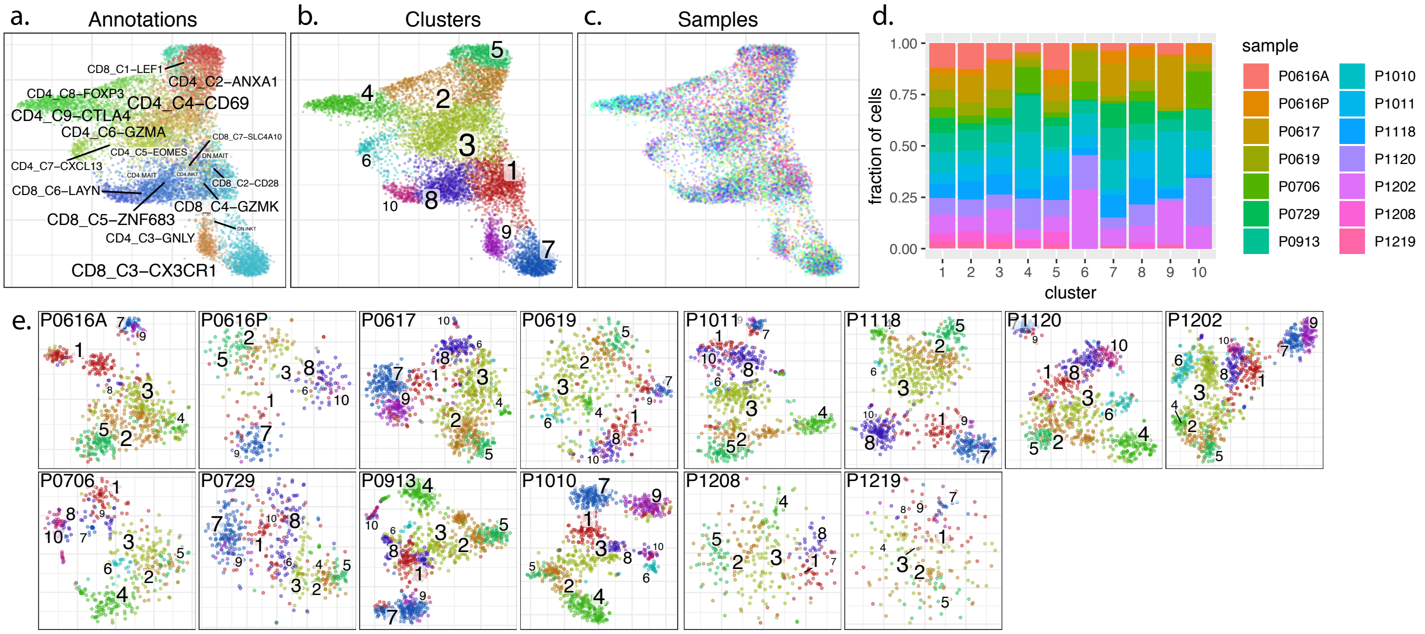
†Correspondence should be addressed to peter_kharchenko@hms.harvard.edu



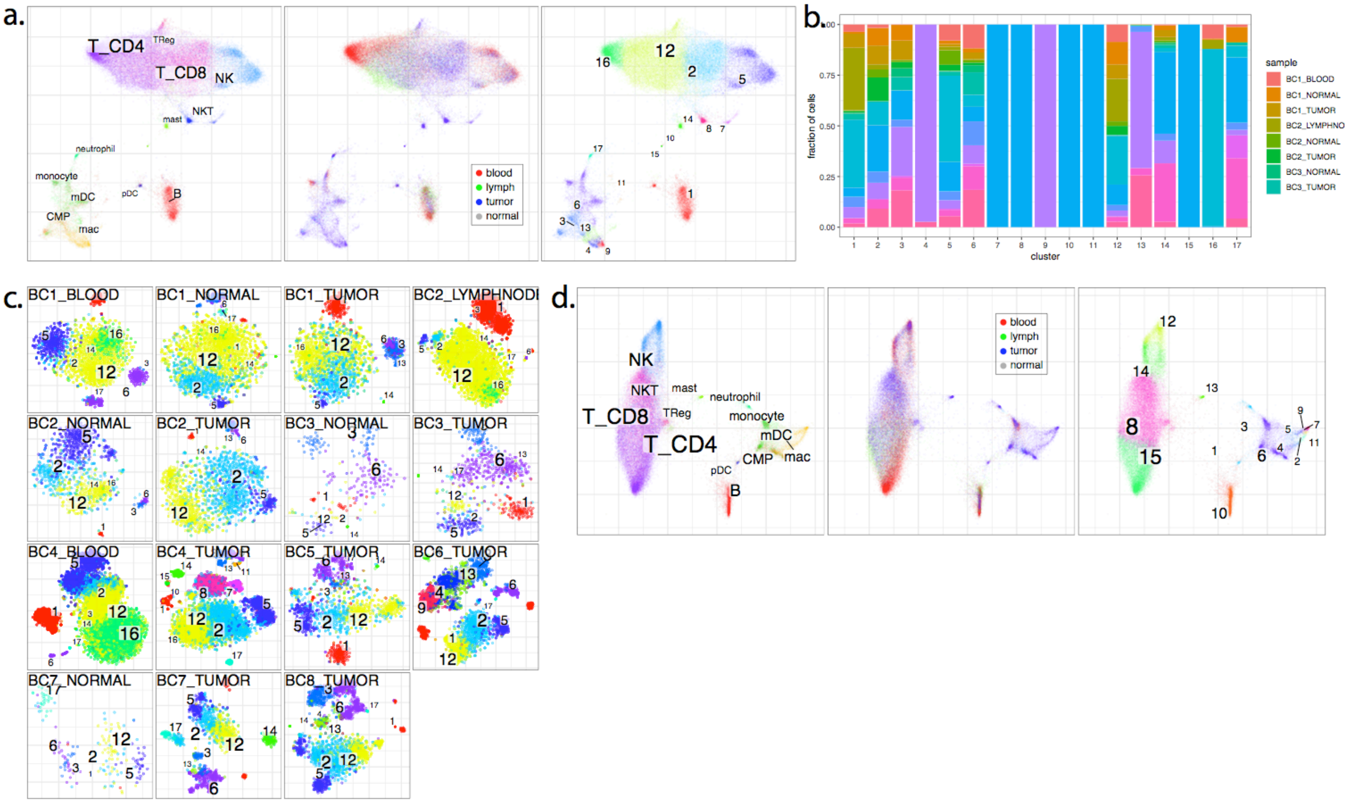
Supplementary Figure 1. Additional analysis of HCA BM+CB 3K dataset. a. Joint graph default embedding, as shown in Figure 1b of the main manuscript. **b.** Expression of select marker genes. **c.** An alternative embedding of the same joint graph as shown in panel a, performed using node2vec. **d.** Distribution of samples within the graph2vis embedding. **e.** Adjusted rand index is shown for the three methods as a function of the p_{mix} probability, with high probability pushing expression of each cell in the dataset closer to dataset-wide average expression profile. **f.** Fraction of single cells correctly classified in a "sensitivity to individual cells" assay (see Methods), where only a single cell was left in a randomly chosen cluster in a random dataset, and the ability of different methods to classify it correctly was measured. n=350 cells tested; Whiskers show the 95% confidence intervals of the binomial proportion. **g.** Comparison of adjusted Rand index (y axis) performance in the heterogeneity benchmark (see Methods, Figure 1f of the main manuscript) for different values of the neighborhood size *k* is shown. Low sensitivity to variation in *k* is observed. **h.** Comparison of different spaces and mappings for pairwise dataset alignments. **i.** A largeVis embedding of the joint graph constructed using nearest neighbor mapping. e,g,h: smoothed estimate of the mean is shown. Shading shows the 95% confidence band of the mean. n=10 random samples per point were used.



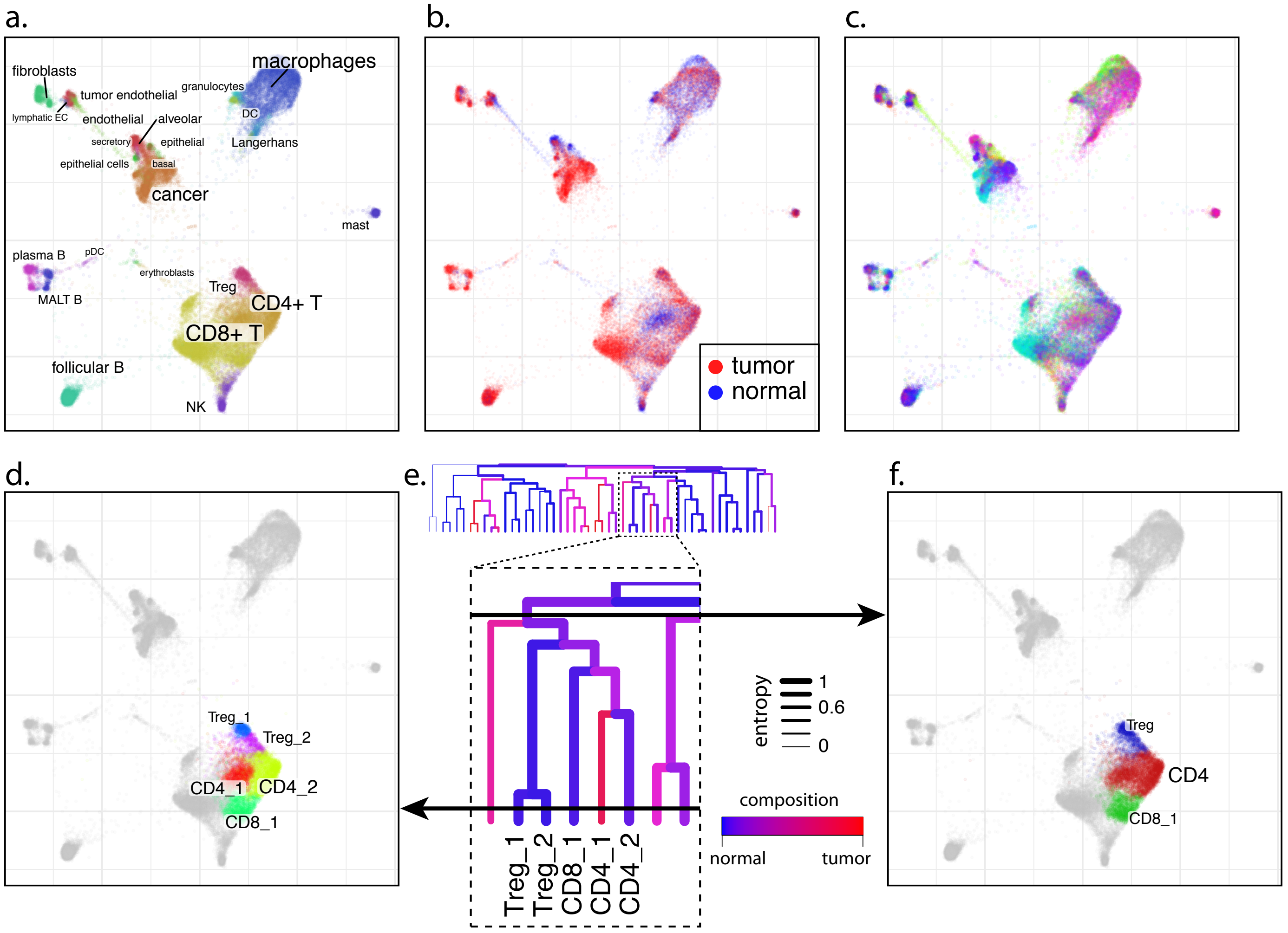
Supplementary Figure 2. Re-analysis of Puram *et al.* dataset on head and neck cancer. a. Joint graph embedding, labeled by the annotations taken from the original publication. **b.** Clusters, as determined by conos on the joint graph. **c.** Mixing of samples is illustrated with different colors. **d.** Sample composition of each conos joint cluster is shown. **e.** t-SNE embeddings of the individual samples that were fed into to the conos analysis, colored and labeled according to conos joint clusters (panel b).



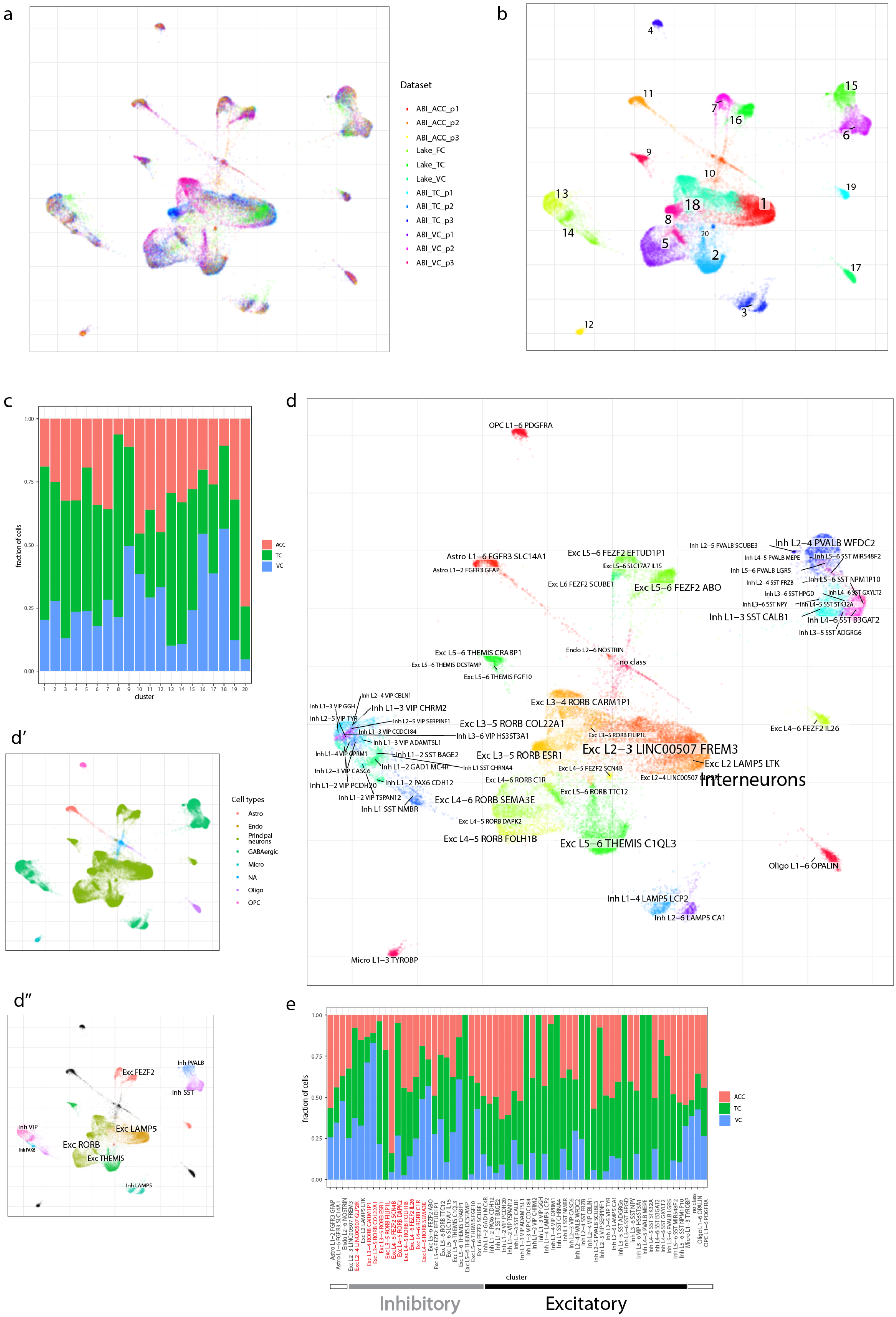
Supplementary Figure 3. Re-analysis of Guo *et al.* dataset on non-small cell lung cancer. a. Joint graph embedding, labeled by the annotations taken from the original publication. **b.** Clusters, as determined by conos on the joint graph. **c.** Mixing of samples is illustrated with different colors. **d.** Sample composition of each conos joint cluster is shown. **e.** t-SNE embeddings of the individual samples that were fed into to the conos analysis, colored and labeled according to conos joint clusters (panel b).



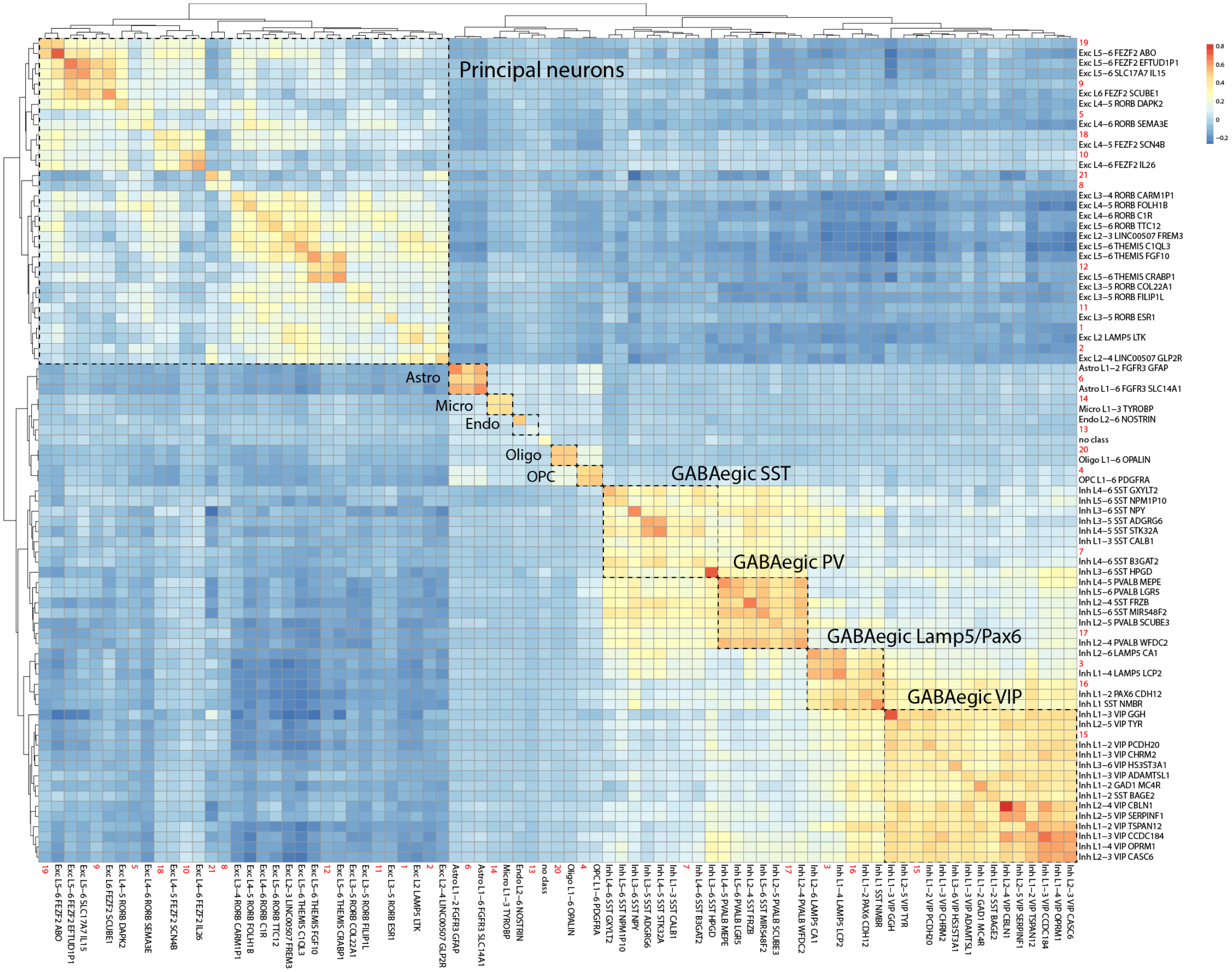
Supplementary Figure 4. Re-analysis of Azizi *et al.* dataset on breast cancer. a. Re-analysis starting with independent patient+tissue slices of the dataset. Joint graph embedding is shown in the left panel, labeled by the annotations taken from the original publication. The annotations have been simplified to collapse patient-specific subpopulations and minor subpopulations. The central panel shows tissue distribution. The right panel shows conos joint clusters. **b.** Sample composition of each conos joint cluster is shown. **c.** t-SNE embeddings of the individual samples that were fed into to the conos analysis shown in a-b, colored and labeled according to conos joint clusters (right panel of a). **d.** Re-analysis starting with 53 independent patient+tissue+replicate slices of the dataset, reveals the same subpopulation structure despite large initial fragmentation of the dataset.



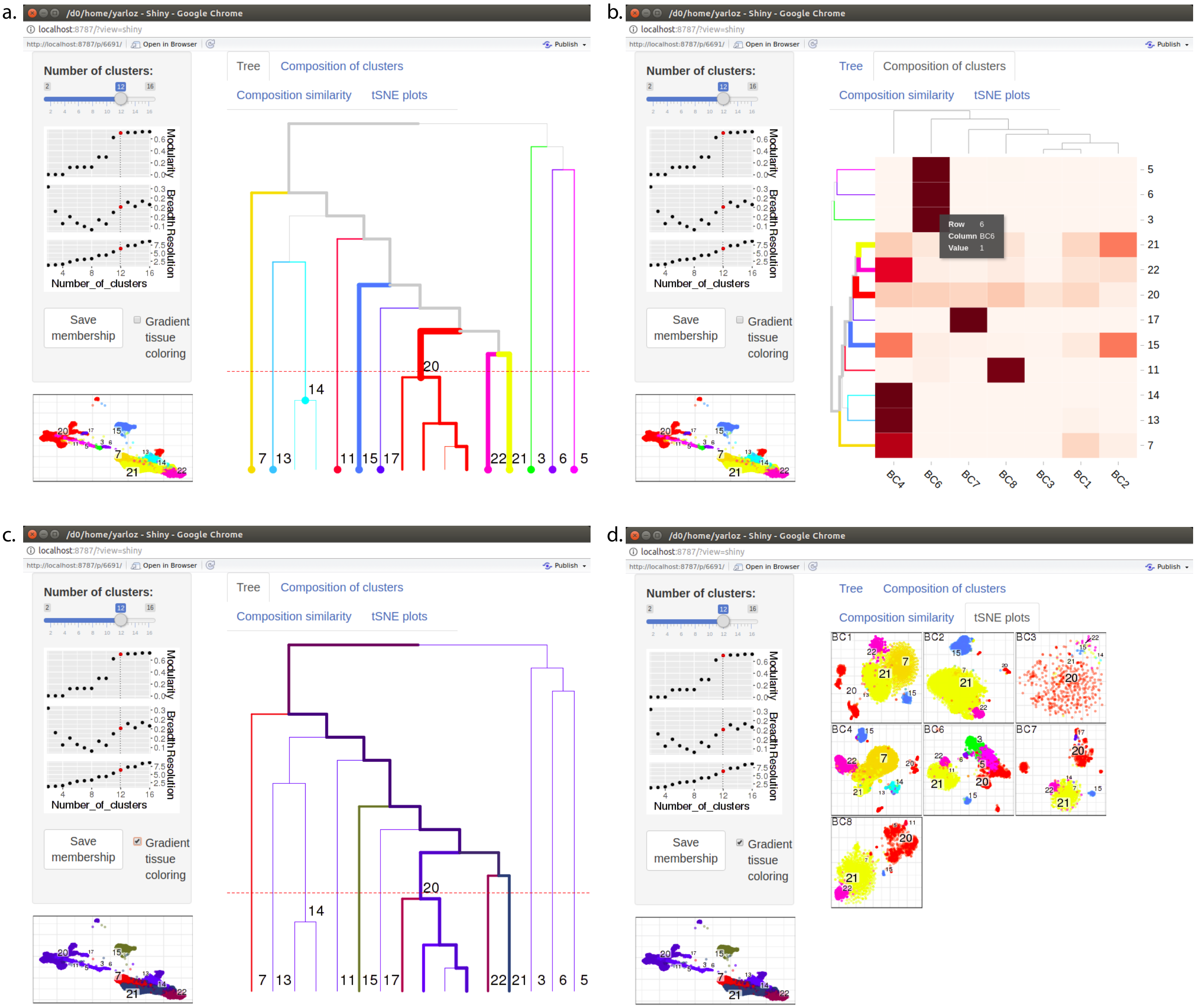
Supplementary Figure 5. Re-analysis of Lambrechts *et al.* dataset on lung cancer. a. Joint graph embedding, labeled by the annotations taken from the original publication. **b.** Tissue distribution is shown on the joint graph embedding. **c.** Sample distribution is shown on the joint graph embedding. **d-f.** Similar to Figure 2c-e of the main manuscript, the panels show T cell subpopulations resulting from two different levels of the joint graph community structure (dendrogram). Lower cut results in more granular T cell subpopulations (panel d), however also results in clusters that show high tissue specificity (e.g. CDr_1 is composed almost entirely of tumor cells). At the same time a higher cut results is less granular clusters that more evenly mix tissues and samples.


a
Dataset
ABI_ACC_p1
ABI_ACC_p2
ABI_ACC_p3
Lake_FC
Lake_TC
Lake_VC
ABI_TC_p1
ABI_TC_p2
ABI_TC_p3
ABI_VC_p1
ABI_VC_p2
ABI_VC_p3
b
4
11
7
16
15
6
9
10
19
13
18
1
8
14
20
5
2
17
3
12
c
1.00
0.75
0.50
0.25
0.00
fraction of cells
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
cluster
ACC
TC
VC
d
OPC L1–6 PDGFRA
Inh L2–5 PVALB SCUBE3
Inh L2–4 PVALB WFDC2
Inh L4–5 PVALB MEPE
Inh L5–6 SST MIR548F2
Astro L1–6 FGFR3 SLC14A1
Exc L5–6 FEZF2 EFTUD1P1
Astro L1–2 FGFR3 GFAP
Exc L5–6 SLC17A7 IL15
Inh L5–6 PVALB LGR5
Inh L5–6 SST NPM1P10
Exc L6 FEZF2 SCUBE1
Exc L5–6 FEZF2 ABO
Inh L2–4 SST FRZB
Inh L3–6 SST HPGD
Inh L4–6 SST GXYLT2
Inh L3–6 SST NPY
Inh L4–5 SST STK32A
Inh L1–3 SST CALB1
Inh L4–6 SST B3GAT2
Inh L3–5 SST ADGRG6
Exc L5–6 THEMIS CRABP1
Endo L2–6 NOSTRIN
Exc L5–6 THEMIS DCSTAMP
no class
Exc L5–6 THEMIS FGF10
Inh L2–4 VIP CBLN1
Exc L3–4 RORB CARM1P1
Inh L1–3 VIP GGH
Inh L1–3 VIP CHRM2
Inh L2–5 VIP TYR
Inh L2–5 VIP SERPINF1
Exc L3–5 RORB COL22A1
Exc L4–6 FEZF2 IL26
Inh L1–3 VIP CCDC184
Inh L3–6 VIP HS3ST3A1
Exc L3–5 RORB FILIP1L
Inh L1–3 VIP ADAMTSL1
Exc L2–3 LINC00507 FREM3
Inh L1–4 VIP OPRM1
Inh L1–2 SST BAGE2
Exc L3–5 RORB ESR1
Exc L2 LAMP5 LTK
Inh L1–2 GAD1 MC4R
Inh L1 SST CHRNA4
Inh L2–3 VIP CASC6
Exc L4–5 FEZF2 SCN4B
Exc L2–4 LINC00507 GLP2R
Interneurons
Inh L1–2 VIP PCDH20
Inh L1–2 PAX6 CDH12
Exc L4–6 RORB C1R
Exc L5–6 RORB TTC12
Inh L1–2 VIP TSPAN12
Exc L4–6 RORB SEMA3E
Inh L1 SST NMBR
Exc L4–5 RORB DAPK2
Exc L4–5 RORB FOLH1B
Exc L5–6 THEMIS C1QL3
Oligo L1–6 OPALIN
Inh L1–4 LAMP5 LCP2
Inh L2–6 LAMP5 CA1
Micro L1–3 TYROBP
d′
Cell types
Astro
Endo
Principal neurons
GABAergic
Micro
NA
Oligo
OPC
d″
Inh PVALB
Exc FEZF2
Inh SST
Inh VIP
Exc LAMP5
Exc RORB
Inh PAX6
Exc THEMIS
Inh LAMP5
e
1.00
0.75
0.50
0.25
0.00
fraction of cells
Astro L1–2 FGFR3 GFAP
Astro L1–6 FGFR3 SLC14A1
Endo L2–6 NOSTRIN
Exc L2–3 LINC00507 FREM3
Exc L2–4 LINC00507 GLP2R
Exc L2 LAMP5 LTK
Exc L3–4 RORB CARM1P1
Exc L3–5 RORB COL22A1
Exc L3–5 RORB ESR1
Exc L3–5 RORB FILIP1L
Exc L4–5 FEZF2 SCN4B
Exc L4–5 RORB DAPK2
Exc L4–5 RORB FOLH1B
Exc L4–6 FEZF2 IL26
Exc L4–6 RORB C1R
Exc L4–6 RORB SEMA3E
Exc L5–6 FEZF2 ABO
Exc L5–6 FEZF2 EFTUD1P1
Exc L5–6 RORB TTC12
Exc L5–6 SLC17A7 IL15
Exc L5–6 THEMIS C1QL3
Exc L5–6 THEMIS CRABP1
Exc L5–6 THEMIS DCSTAMP
Exc L5–6 THEMIS FGF10
Exc L6 FEZF2 SCUBE1
Inh L1–2 GAD1 MC4R
Inh L1–2 PAX6 CDH12
Inh L1–2 SST BAGE2
Inh L1–2 VIP PCDH20
Inh L1–2 VIP TSPAN12
Inh L1–3 SST CALB1
Inh L1–3 VIP ADAMTSL1
Inh L1–3 VIP CCDC184
Inh L1–3 VIP CHRM2
Inh L1–3 VIP GGH
Inh L1–4 LAMP5 LCP2
Inh L1–4 VIP OPRM1
Inh L1 SST CHRNA4
Inh L1 SST NMBR
Inh L2–3 VIP CASC6
Inh L2–4 PVALB WFDC2
Inh L2–4 SST FRZB
Inh L2–4 VIP CBLN1
Inh L2–5 PVALB SCUBE3
Inh L2–5 VIP SERPINF1
Inh L2–5 VIP TYR
Inh L2–6 LAMP5 CA1
Inh L3–5 SST ADGRG6
Inh L3–6 SST HPGD
Inh L3–6 SST NPY
Inh L3–6 VIP HS3ST3A1
Inh L4–5 PVALB MEPE
Inh L4–5 SST STK32A
Inh L4–6 SST B3GAT2
Inh L4–6 SST GXYLT2
Inh L5–6 PVALB LGR5
Inh L5–6 SST MIR548F2
Inh L5–6 SST NPM1P10
Micro L1–3 TYROBP
no class
Oligo L1–6 OPALIN
OPC L1–6 PDGFRA
cluster
ACC
TC
VC
Inhibitory
Excitatory

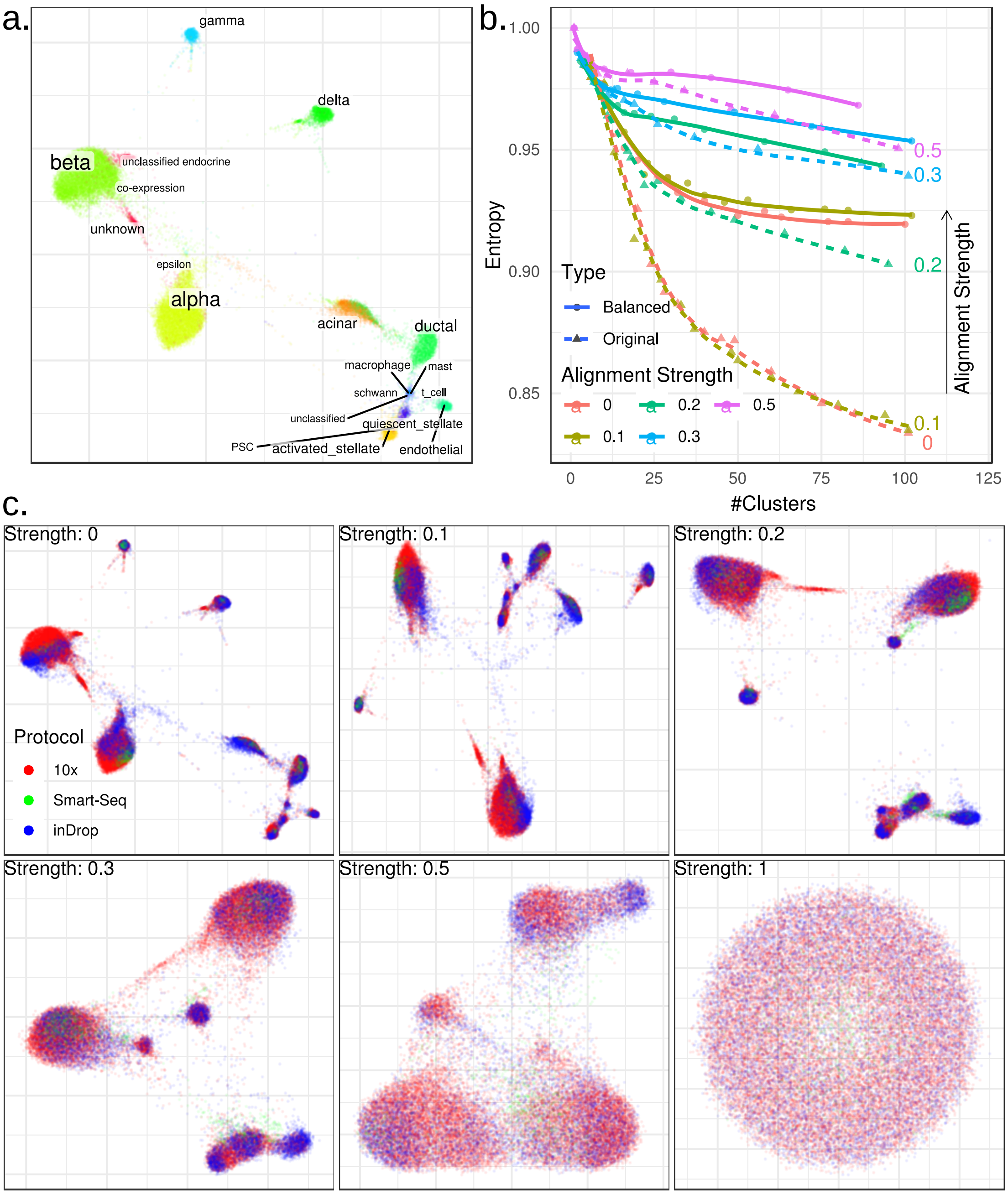
Supplementary Figure 6. Re-analysis of human cortex datasets. a. Conos joint graph embedding labeled by dataset. The following datasets were clustered: temporal (TC), visual (VC) and frontal cortex (FC) from Lake *et al.* (2016) Science; Allen Brain Institute datasets for temporal cortex (TC) from Hodge *et al.* (2018) bioRxiv, for anterior cingulate (ACC), and visual cortex (VC). Note that single cell transcriptomics data for each cortical area from Allen Brain Institute source consists of three datasets each obtained from a different patient (p1, p2 and p3). **b.** Cells labeled by Conos joint clusters (multilevel.community clusters were used). **c.** Cell composition of each Conos cluster based on the cortical area, *i.e.* temporal, visual and anterior cingulate cortices. Note that due to smaller sample size of Lake *et al.*, only cells derived from Allen Brain Institute datasets are shown. **d.** Conos clustering annotated by subtypes described for the temporal cortex data in Hodge *et al.* The annotation was propagated from the Hodge *et al*. temporal cortex dataset to all the other datasets. The same embedding is labeled based on major neuronal and non-neuronal cell types (**d'**) and major classes of principal (Exc) and GABAergic (Inh) neurons as in Hodge *et al*. annotation (**d''**). **e.** Normalized fraction of cells in each subtype from different cortical areas. Note high diversity of layer 4 principal neurons (subtypes are labeled in red) that might be due to differences in cortical laminar architecture*, i.e*. ACC has agranular or dysgranule structure, while TC and VC are eulaminate. Black, grey and white bars below the graph label GABAergic neurons, principal neurons and non-neuronal cells.



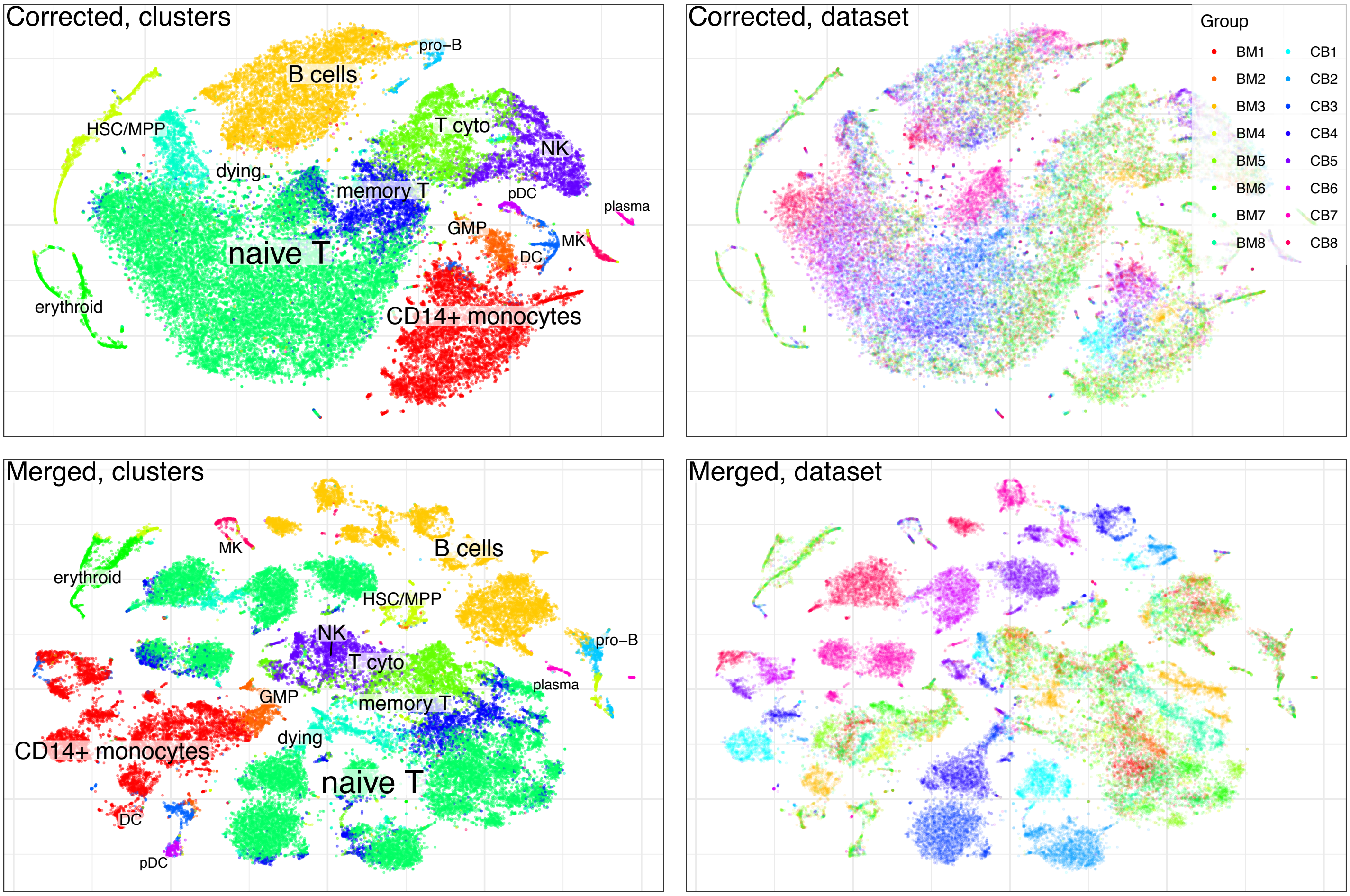
Supplementary Figure 7. Correlation of gene expression for human cortical single cell transcriptomes between cell clusters determined by Conos and cortical cell subtypes annotated in Hodge *et al.* bioRxiv 2018. Conos clusters are labeled in red. Note generally good correlation of Conos clustering with subtypes clustered by Hodge *et al.* that allows to annotate Conos clusters, *e.g.* major subtypes of GABAergic interneurons in Conos embedding are clusters 15 (VIP-expressing), 16 and 3 (Lamp5/Pax6-expressing), 17 (PV-expressing) and 7 (SST-expressing). Pearson linear correlation was used.



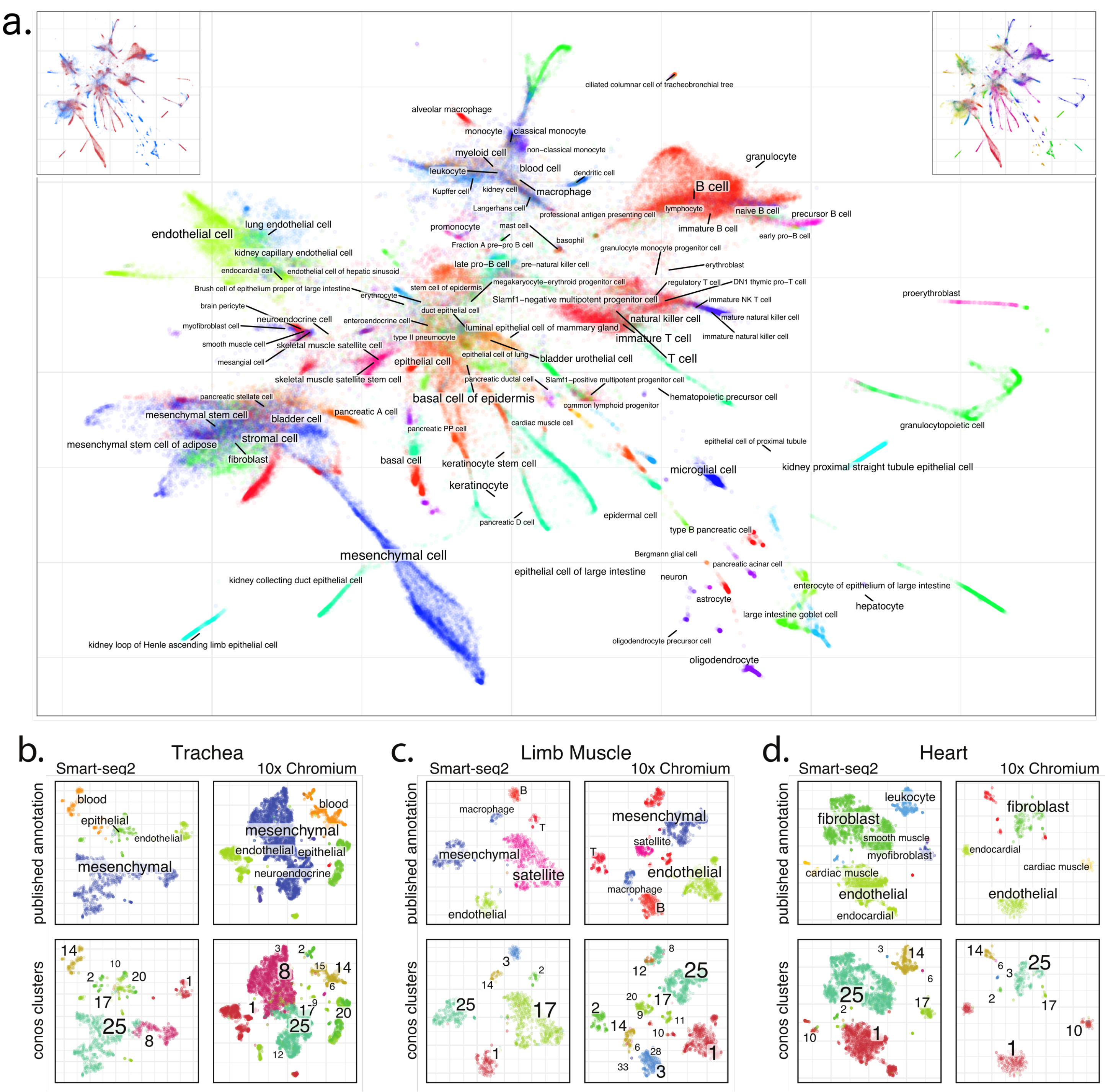
Supplementary Figure 8. Interactive application for navigating community hierarchies. Several screens of the interactive (Shiny) application for choosing appropriate cuts in the hierarchical community structures are shown. **a.** Initial screen, allowing user to select the number of top clusters, visualize tree and joint graph embedding colored by the resulting clusters. **b.** Heatmap visualization of the interactions between samples and clusters. Such heatmaps allow users to pick up on separation of both clusters and samples based on their composition in the joint clustering analysis. **c.** Community hierarchy clustered by tissue composition. **d.** Individual t-SNE sample embeddings, colored by joint clusters.

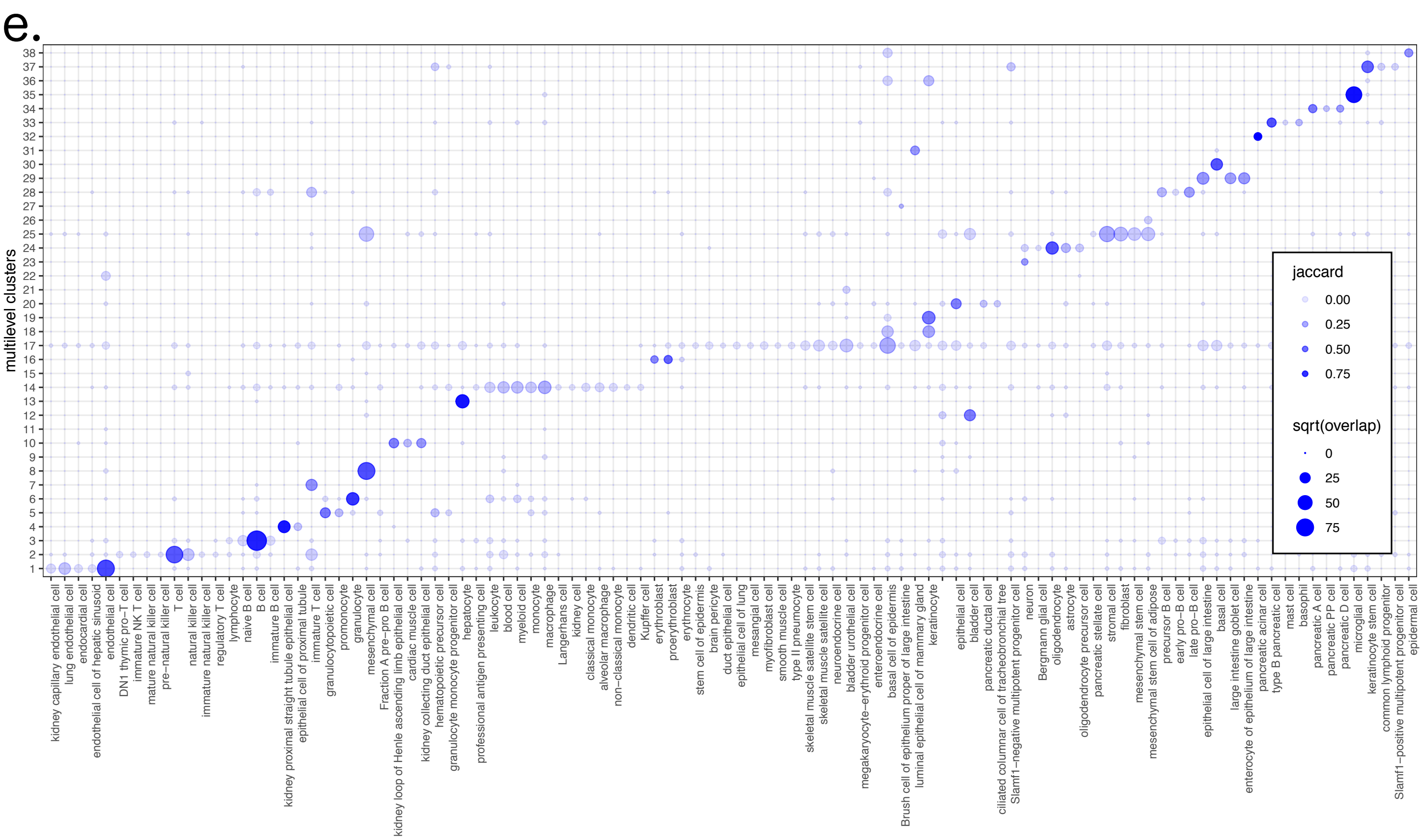


Supplementary Figure 9. Improvement of alignment quality over protocols with edge rebalancing and increase of alignment strength. a. Alignment of three human pancreas islet datasets (Datasets 10 in the “Data sources” listing), obtained using different protocols (10x Chromium, inDrops and Smart-seq2). The distribution of the cells associated with different platforms in the default Conos configuration is show in the top left corner of panel c (Strength 0). **c.** Increasing alignment strength α from 0.0 to 1.0, we obtained 6 graphs and visualized them using largeVis embeddings, coloring cells by the protocol. The value of the alignment strength parameter α is shown in the top-left corner of each plot. The parameter α implements a trade-off between sample mixing and cluster resolution. **b.** To validate quality of each graph we estimated normalized relative entropy for Leiden clustering, uniformly varying its resolution parameter to obtain 15 different clusterings, which have between 1 and 100 clusters each. The same procedure was repeated with and without edge weight rebalancing on each graph, which is designed to further improve inter-platform alignment. The results are shown in panel c., where x-axis corresponds to number of clusters, y-axis represents entropy, colors show alignment strength and the line/marker style distinguishes results with and without edge weight balancing. While higher resolution clustering will always yield lower entropy, increasing the alignment strength parameter α at a given cluster resolution improves mixing of the cells from different platforms. Moreover, for the same alignment level, edge balancing leads to significant increase in entropy.

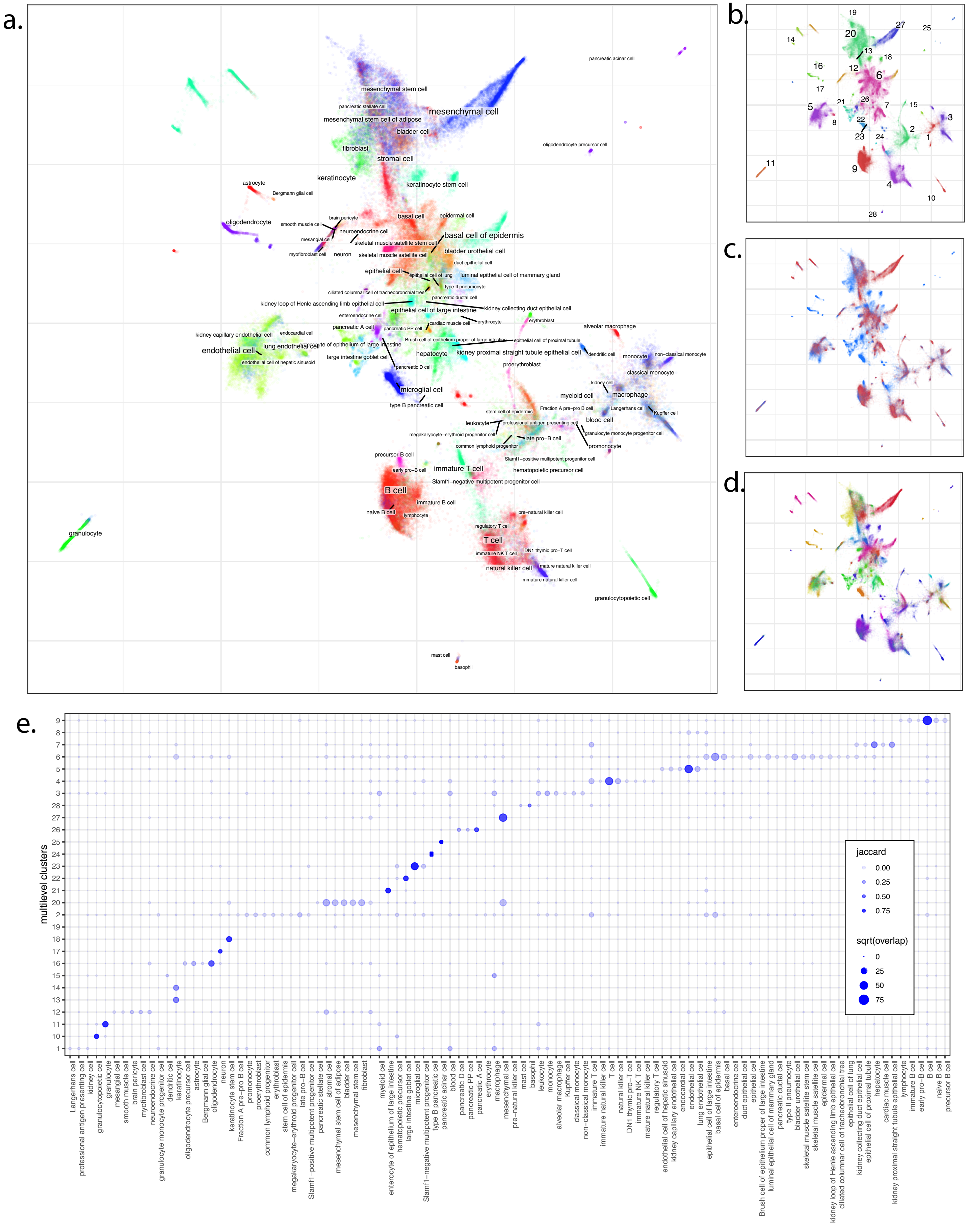


Supplementary Figure 10. Estimation of common expression space by diffusion on a joint graph. The top panels show t-SNE embeddings of the HCA BM+CB 3k dataset after "correction" of the expression values through graph diffusion – a process that estimates "common" or "corrected" expression coordinates for all of the datasets. The left top plot shows major subpopulations (as in Figure 1b of the main manuscript), and the right panel shows distribution of the different datasets. The bottom row shows equivalent embeddings obtained by simply joining molecular count matrices of the different datasets without any additional corrections. Such processing leaves pronounced patient/batch effect.

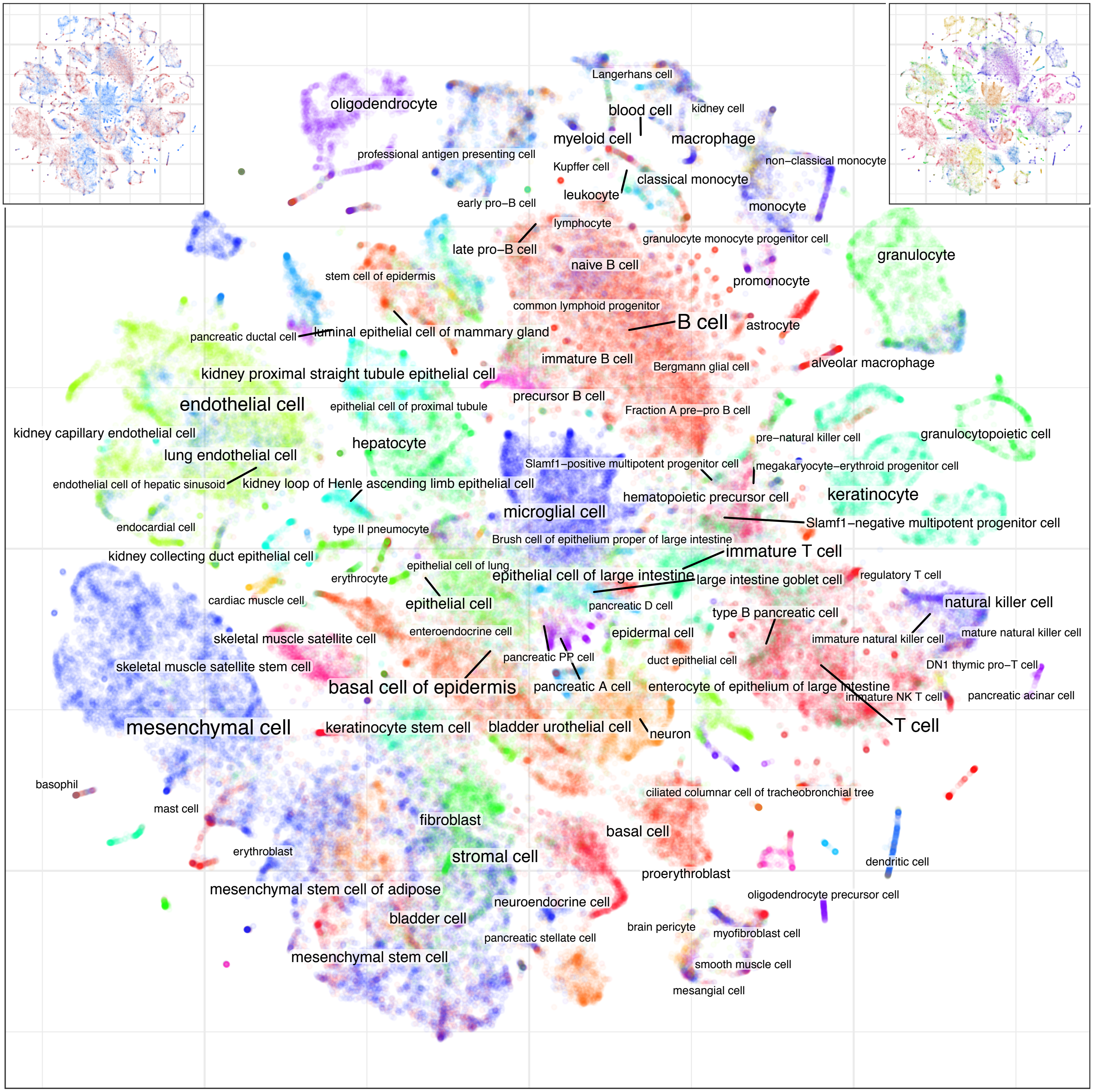

a.
ciliated columnar cell of tracheobronchial tree
alveolar macrophage
monocyte
classical monocyte
non−classical monocyte
myeloid cell
granulocyte
blood cell
leukocyte
dendritic cell
kidney cell
Kupffer cell
macrophage
B cell
Langerhans cell
lymphocyte
naive B cell
precursor B cell
professional antigen presenting cell
immature B cell
promonocyte
mast cell
early pro−B cell
basophil
Fraction A pre−pro B cell
granulocyte monocyte progenitor cell
endothelial cell
lung endothelial cell
kidney capillary endothelial cell
late pro−B cell
pre−natural killer cell
erythroblast
endocardial cell
endothelial cell of hepatic sinusoid
megakaryocyte−erythroid progenitor cell
regulatory T cell
DN1 thymic pro−T cell
Brush cell of epithelium proper of large intestine
stem cell of epidermis
erythrocyte
Slamf1−negative multipotent progenitor cell
immature NK T cell
proerythroblast
brain pericyte
duct epithelial cell
natural killer cell
mature natural killer cell
neuroendocrine cell
myofibroblast cell
enteroendocrine cell
luminal epithelial cell of mammary gland
type II pneumocyte
immature T cell
immature natural killer cell
smooth muscle cell
skeletal muscle satellite cell
mesangial cell
epithelial cell of lung
bladder urothelial cell
epithelial cell
T cell
skeletal muscle satellite stem cell
pancreatic ductal cell
Slamf1−positive multipotent progenitor cell
pancreatic stellate cell
basal cell of epidermis
hematopoietic precursor cell
common lymphoid progenitor
mesenchymal stem cell
pancreatic A cell
bladder cell
cardiac muscle cell
granulocytopoietic cell
pancreatic PP cell
mesenchymal stem cell of adipose
stromal cell
epithelial cell of proximal tubule
fibroblast
basal cell
keratinocyte stem cell
microglial cell
kidney proximal straight tubule epithelial cell
keratinocyte
epidermal cell
pancreatic D cell
type B pancreatic cell
mesenchymal cell
Bergmann glial cell
pancreatic acinar cell
epithelial cell of large intestine
neuron
kidney collecting duct epithelial cell
enterocyte of epithelium of large intestine
astrocyte
hepatocyte
large intestine goblet cell
oligodendrocyte precursor cell
kidney loop of Henle ascending limb epithelial cell
oligodendrocyte
b.
Trachea
Smart-seq2
10x Chromium
published annotation
blood
epithelial
endothelial
mesenchymal
blood
mesenchymal
endothelial
epithelial
neuroendocrine
conos clusters
14
10
2
20
1
17
25
8
3
2
8
15
14
6
1
9
17
25
20
12
c.
Limb Muscle
Smart-seq2
10x Chromium
published annotation
B
macrophage
T
mesenchymal
satellite
endothelial
mesenchymal
satellite
T
endothelial
macrophage
B
conos clusters
3
2
14
25
17
1
8
12
25
20
17
2
9
14
11
10
6
28
33
3
1
d.
Heart
Smart-seq2
10x Chromium
published annotation
leukocyte
fibroblast
smooth muscle
myofibroblast
cardiac muscle
endothelial
endocardial
fibroblast
endocardial
cardiac muscle
endothelial
conos clusters
3
14
6
25
2
17
1
10
14
6
3
25
2
17
10
1



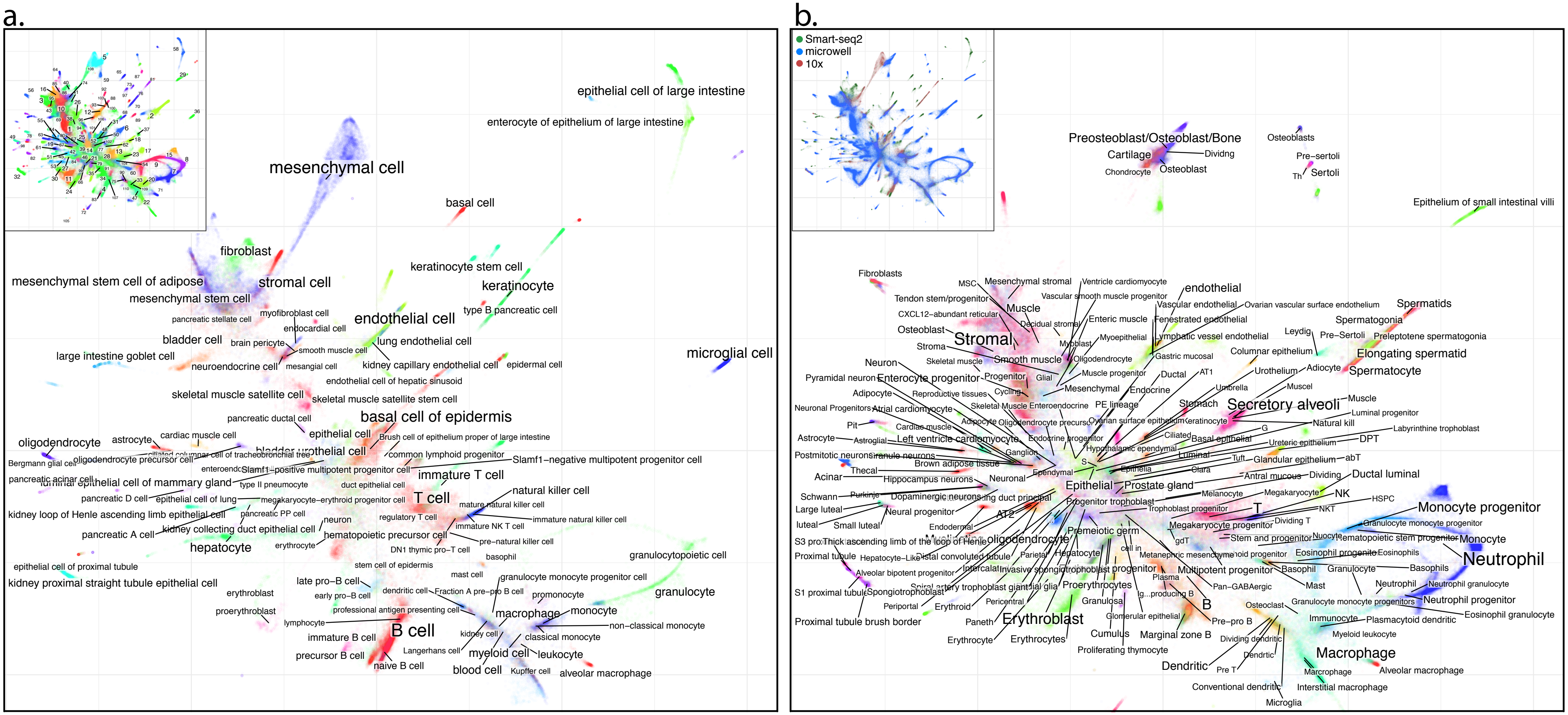
Supplementary Figure 11. Published annotation of Tabula Muris dataset and joint clustering results. a. Published annotations are shown as labels and colors on the joint graph embedding of CPCA space analysis combining 48 separate datasets covering different mouse tissues is shown (100,605 cells). Platform distribution (red – 10x; blue – Smart-seq2) is shown in the top left inset. Distribution of individual samples is shown in the top right inset. **b-d.** Comparison of the joint clusters with the published annotation is shown for three tissues that were measured with both platforms. Joint clustering shows consistency between tissues and platforms, with some clusters giving higher resolution (*e.g.* separation of blood or mesenchymal populations in trachea samples), and others joining related cell types across tissues (*e.g.* fibroblast and part of the mesenchymal population are joined under cluster 4 in l-n). **e.** Details of the correspondence between Conos joint clusters (rows) and Tabula Muris annotation (columns). The size of the circle shows the number of cells, with shading indicating the Jaccard coefficient.


a.
pancreatic acinar cell
mesenchymal stem cell
pancreatic stellate cell
mesenchymal cell
mesenchymal stem cell of adipose
bladder cell
oligodendrocyte precursor cell
fibroblast
stromal cell
keratinocyte
keratinocyte stem cell
astrocyte
Bergmann glial cell
brain pericyte
basal cell
epidermal cell
oligodendrocyte
smooth muscle cell
neuroendocrine cell
mesangial cell
skeletal muscle satellite stem cell
basal cell of epidermis
myofibroblast cell
neuron
skeletal muscle satellite cell
bladder urothelial cell
duct epithelial cell
epithelial cell
epithelial cell of lung
luminal epithelial cell of mammary gland
type II pneumocyte
ciliated columnar cell of tracheobronchial tree
pancreatic ductal cell
kidney loop of Henle ascending limb epithelial cell
kidney collecting duct epithelial cell
epithelial cell of large intestine
enteroendocrine cell
erythrocyte
erythroblast
cardiac muscle cell
pancreatic A cell
pancreatic PP cell
alveolar macrophage
kidney capillary endothelial cell
endocardial cell
Brush cell of epithelium proper of large intestine
epithelial cell of proximal tubule
lung endothelial cell
te of epithelium of large intestine
endothelial cell
hepatocyte
kidney proximal straight tubule epithelial cell
dendritic cell
monocyte
non−classical monocyte
large intestine goblet cell
endothelial cell of hepatic sinusoid
pancreatic D cell
proerythroblast
classical monocyte
kidney cell
microglial cell
myeloid cell
macrophage
type B pancreatic cell
stem cell of epidermis
Fraction A pre−pro B cell
Langerhans cell
Kupffer cell
leukocyte
professional antigen presenting cell
blood cell
megakaryocyte−erythroid progenitor cell
late pro−B cell
granulocyte monocyte progenitor cell
common lymphoid progenitor
promonocyte
Slamf1−positive multipotent progenitor cell
precursor B cell
early pro−B cell
immature T cell
hematopoietic precursor cell
Slamf1−negative multipotent progenitor cell
B cell
immature B cell
naive B cell
lymphocyte
pre−natural killer cell
granulocyte
regulatory T cell
T cell
DN1 thymic pro−T cell
immature NK T cell
natural killer cell
mature natural killer cell
immature natural killer cell
granulocytopoietic cell
mast cell
basophil
b.
19
27
25
20
14
13
18
12
6
16
17
26
21
15
5
7
3
8
22
2
1
23
24
11
9
4
10
28
c.
d.
e.
multilevel clusters
9
8
7
6
5
4
3
28
27
26
25
24
23
22
21
20
2
19
18
17
16
15
14
13
12
11
10
1
jaccard
0.00
0.25
0.50
0.75
sqrt(overlap)
0
25
50
75
Langerhans cell
professional antigen presenting cell
kidney cell
granulocytopoietic cell
granulocyte
mesangial cell
smooth muscle cell
brain pericyte
myofibroblast cell
neuroendocrine cell
granulocyte monocyte progenitor cell
dendritic cell
keratinocyte
oligodendrocyte precursor cell
astrocyte
Bergmann glial cell
oligodendrocyte
neuron
keratinocyte stem cell
Fraction A pre−pro B cell
promonocyte
proerythroblast
common lymphoid progenitor
erythroblast
stem cell of epidermis
megakaryocyte−erythroid progenitor cell
late pro−B cell
Slamf1−positive multipotent progenitor cell
pancreatic stellate cell
stromal cell
mesenchymal stem cell of adipose
bladder cell
mesenchymal stem cell
fibroblast
myeloid cell
enterocyte of epithelium of large intestine
hematopoietic precursor cell
large intestine goblet cell
microglial cell
Slamf1−negative multipotent progenitor cell
type B pancreatic cell
pancreatic acinar cell
blood cell
pancreatic D cell
pancreatic PP cell
pancreatic A cell
erythrocyte
macrophage
mesenchymal cell
pre−natural killer cell
mast cell
basophil
leukocyte
monocyte
alveolar macrophage
Kupffer cell
classical monocyte
non−classical monocyte
immature T cell
immature natural killer cell
T cell
natural killer cell
DN1 thymic pro−T cell
immature NK T cell
mature natural killer cell
regulatory T cell
endothelial cell of hepatic sinusoid
kidney capillary endothelial cell
endocardial cell
endothelial cell
lung endothelial cell
epithelial cell of large intestine
basal cell of epidermis
basal cell
enteroendocrine cell
duct epithelial cell
epithelial cell
Brush cell of epithelium proper of large intestine
luminal epithelial cell of mammary gland
pancreatic ductal cell
type II pneumocyte
bladder urothelial cell
skeletal muscle satellite stem cell
skeletal muscle satellite cell
epidermal cell
kidney loop of Henle ascending limb epithelial cell
ciliated columnar cell of tracheobronchial tree
kidney collecting duct epithelial cell
epithelial cell of lung
epithelial cell of proximal tubule
hepatocyte
cardiac muscle cell
kidney proximal straight tubule epithelial cell
lymphocyte
immature B cell
early pro−B cell
B cell
naive B cell
precursor B cell

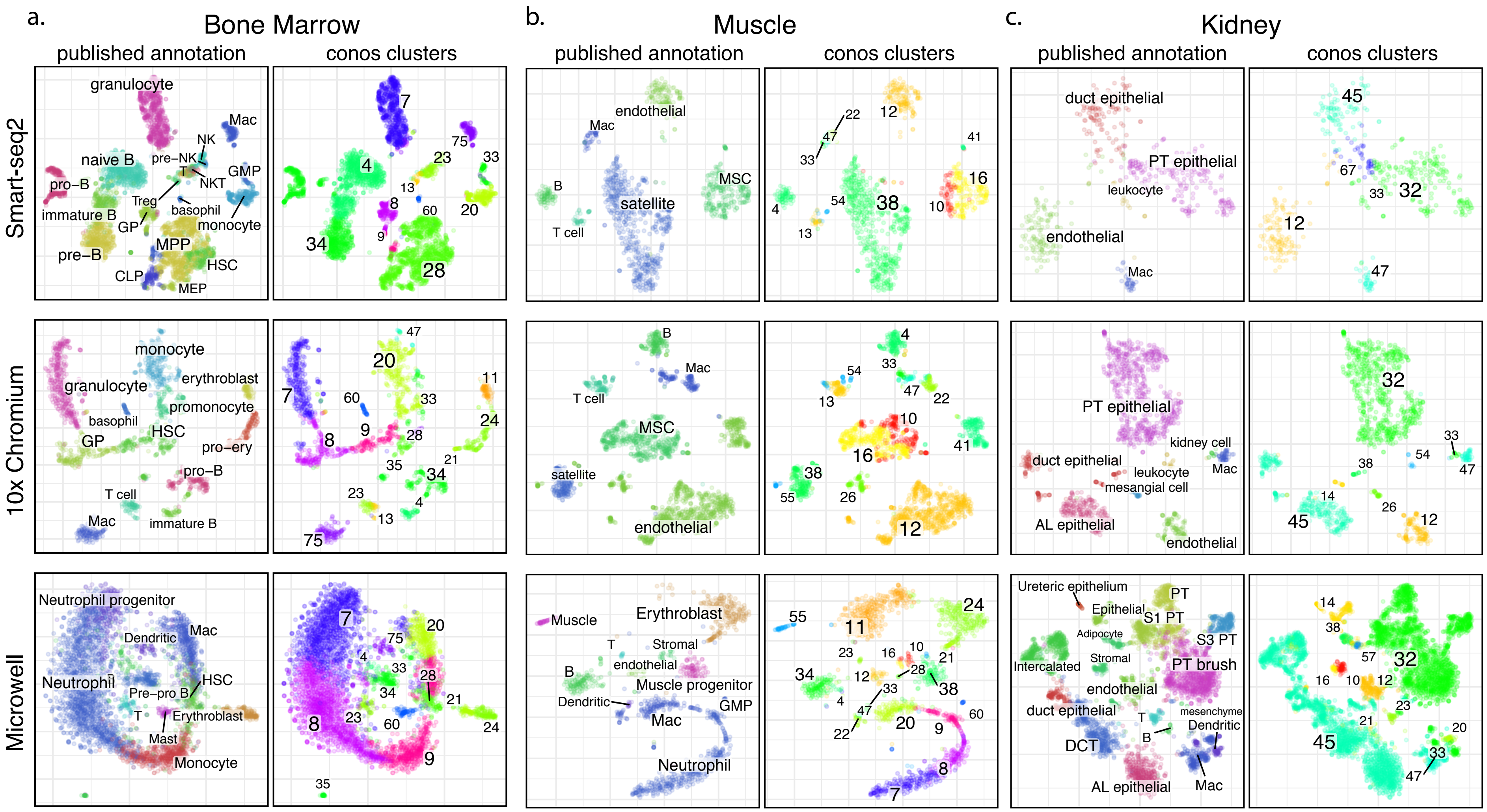
Supplementary Figure 12. Tabula Muris joint analysis using gene space. a. Published annotations are shown as labels and colors on the embedding of a joint graph determined using gene space analysis. **b.** Joint clusters determined by Conos. **c.** Distribution of platforms (red- 10x; blue- Smart-seq2). **d.** Distribution of individual samples. **e.** Correspondence of Conos joint clusters to published annotations.



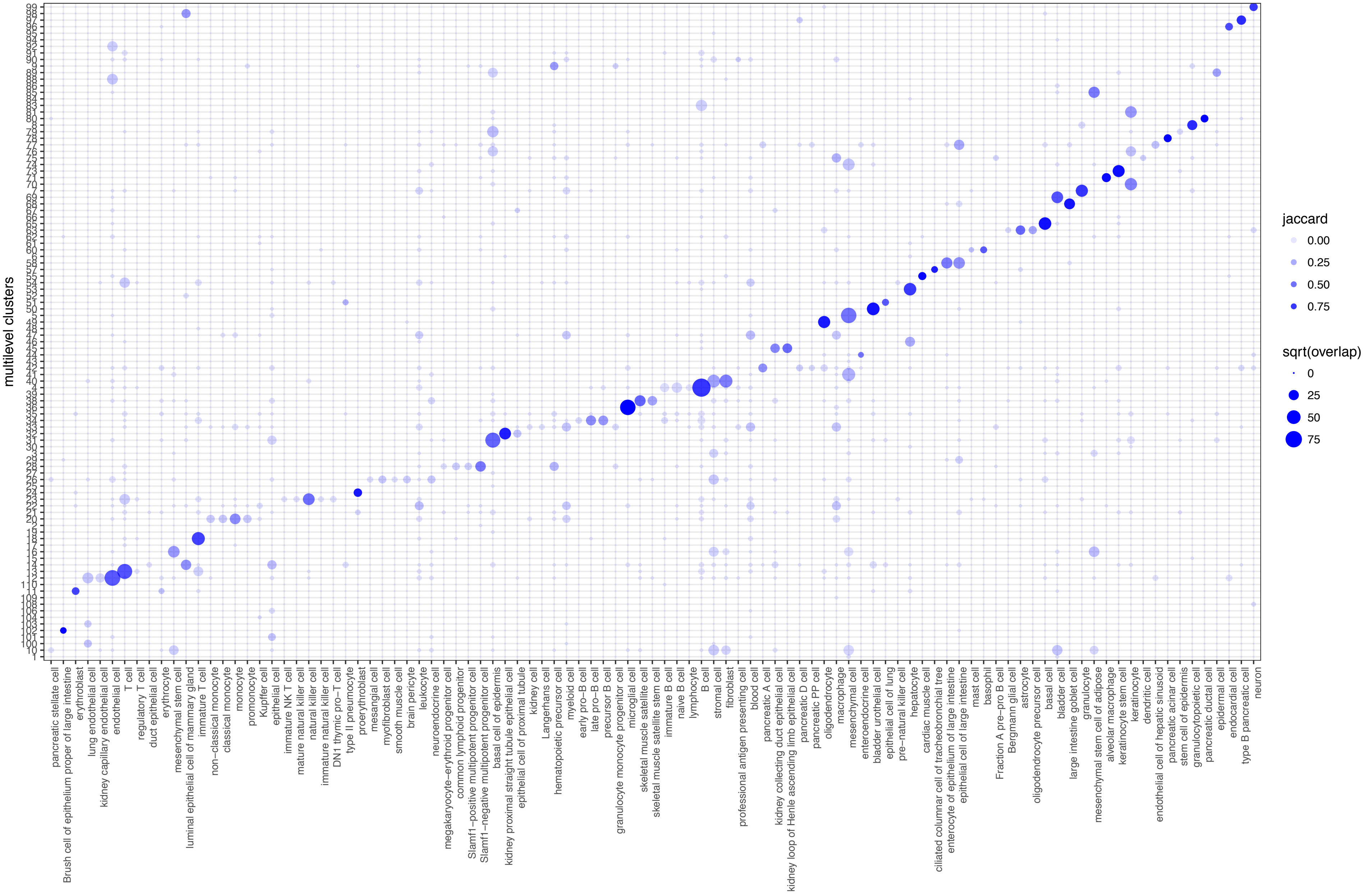
Supplementary Figure 13. Embedding of Tabula Muris joint graph (CPCA space) using graph2vec. Distribution of platforms is shown in the top-left inset (red-10x, blue-Smart-seq2). Distribution of individual samples is shown in the top-right inet.



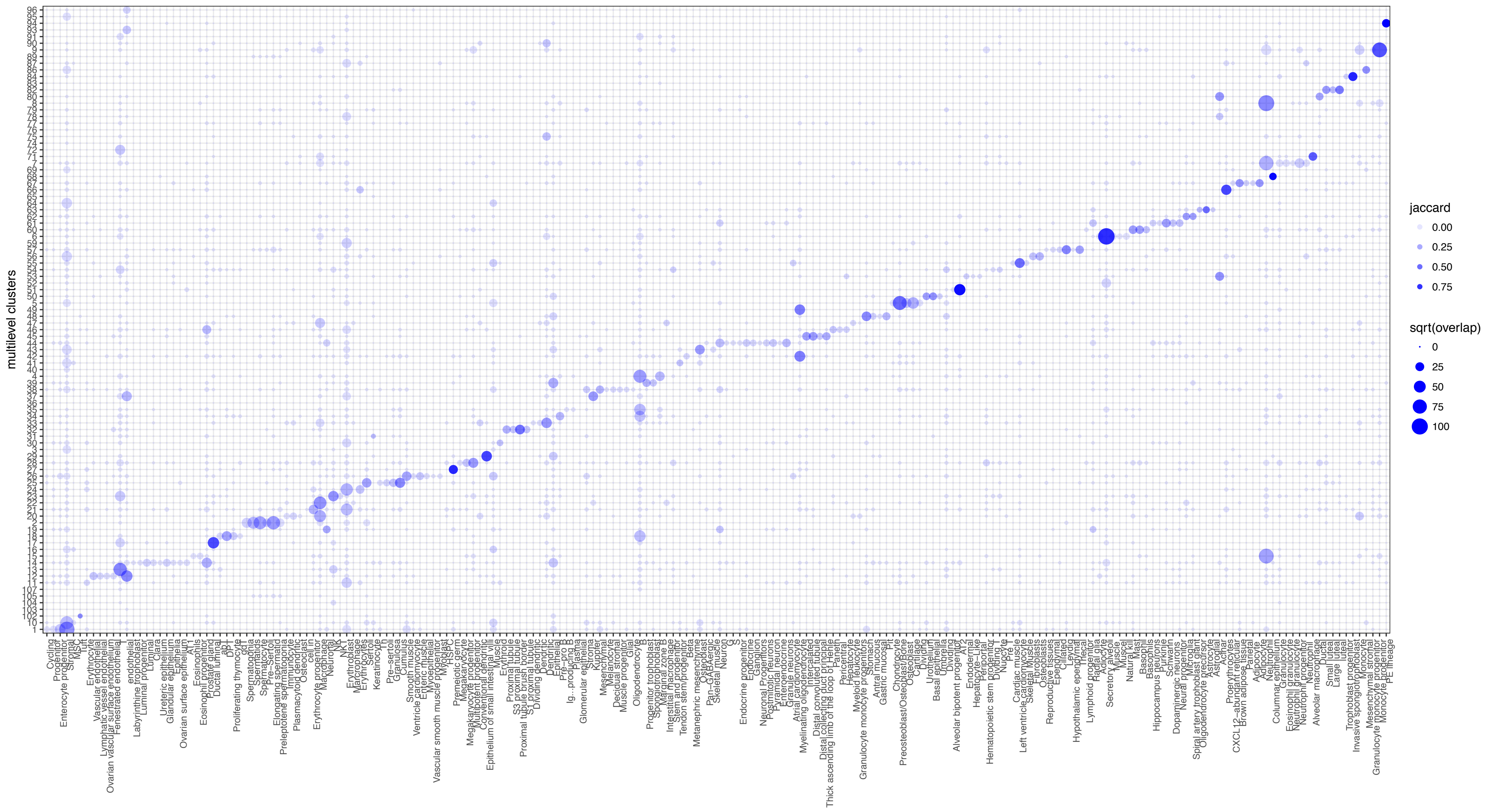
Supplementary Figure 14. Published annotation of two mouse atlases on a joint graph embedding. The analysis combined Tabula Muris and Han *et al.* mouse atlases, joining a total of 173 datasets (>400k+ cells) measured using one of three different platforms. The joint embedding plots show annotations from Tabula Muris (**a.**) and Han *et al.* (**b.**), in each case showing only the subset of cells in the joint embedding from one of the atlases. The top left inset shows overview of complete joint embedding (all 174 datasets), and right inset shows distribution of the platforms within the joint embedding.



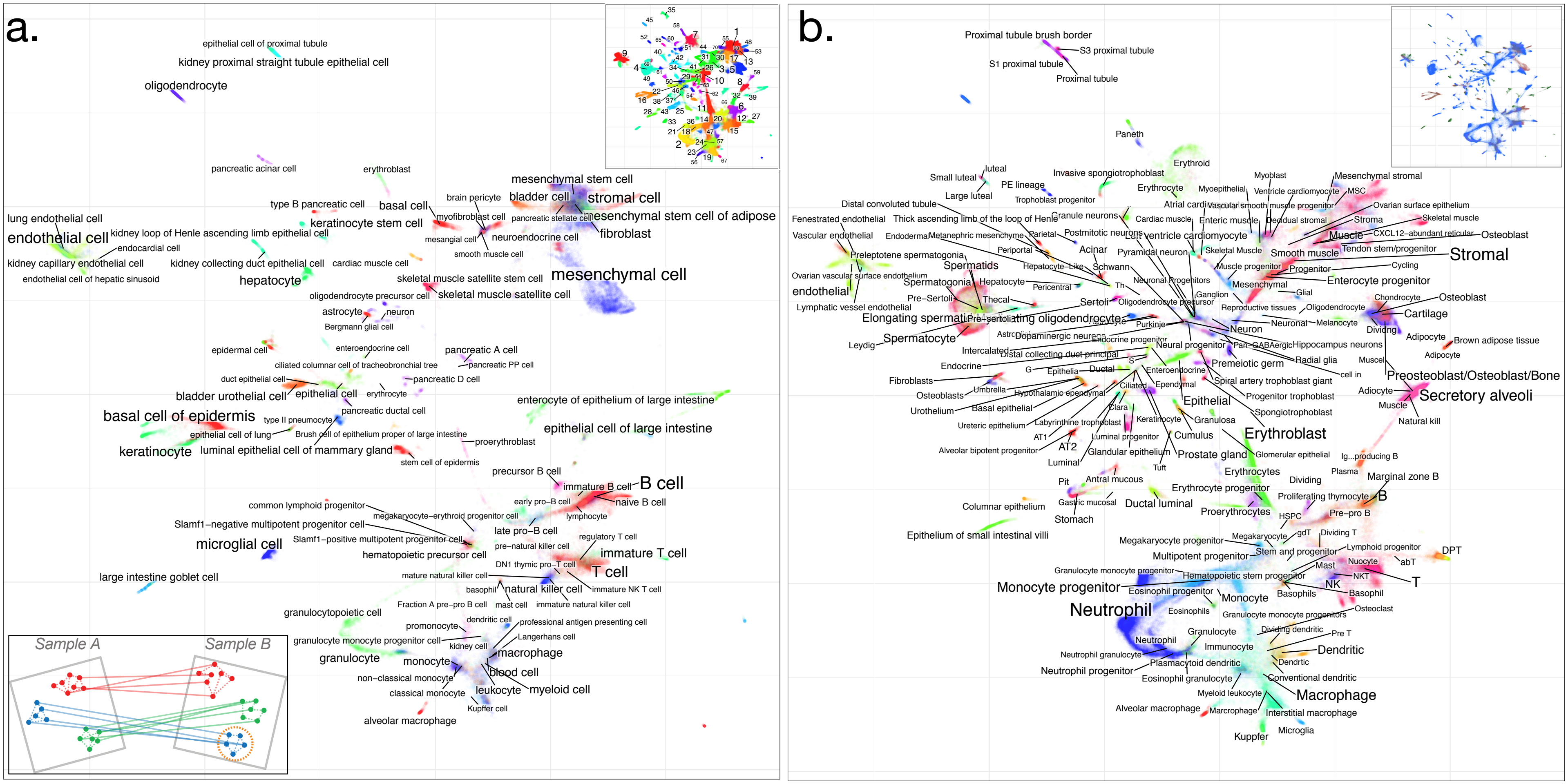
Supplementary Figure 15. Correspondence of published annotations and Conos clusters across mouse atlases. Individual samples from there different tissues (a-c) contained in the Tabula Muris (Smart-seq2 and 10x Chromium rows) and Han et al (Mcirowell row) are shown, together with Conos clusters called on the joint graph combining the two atlases. Joint clustering shows consistency between tissues and platforms.



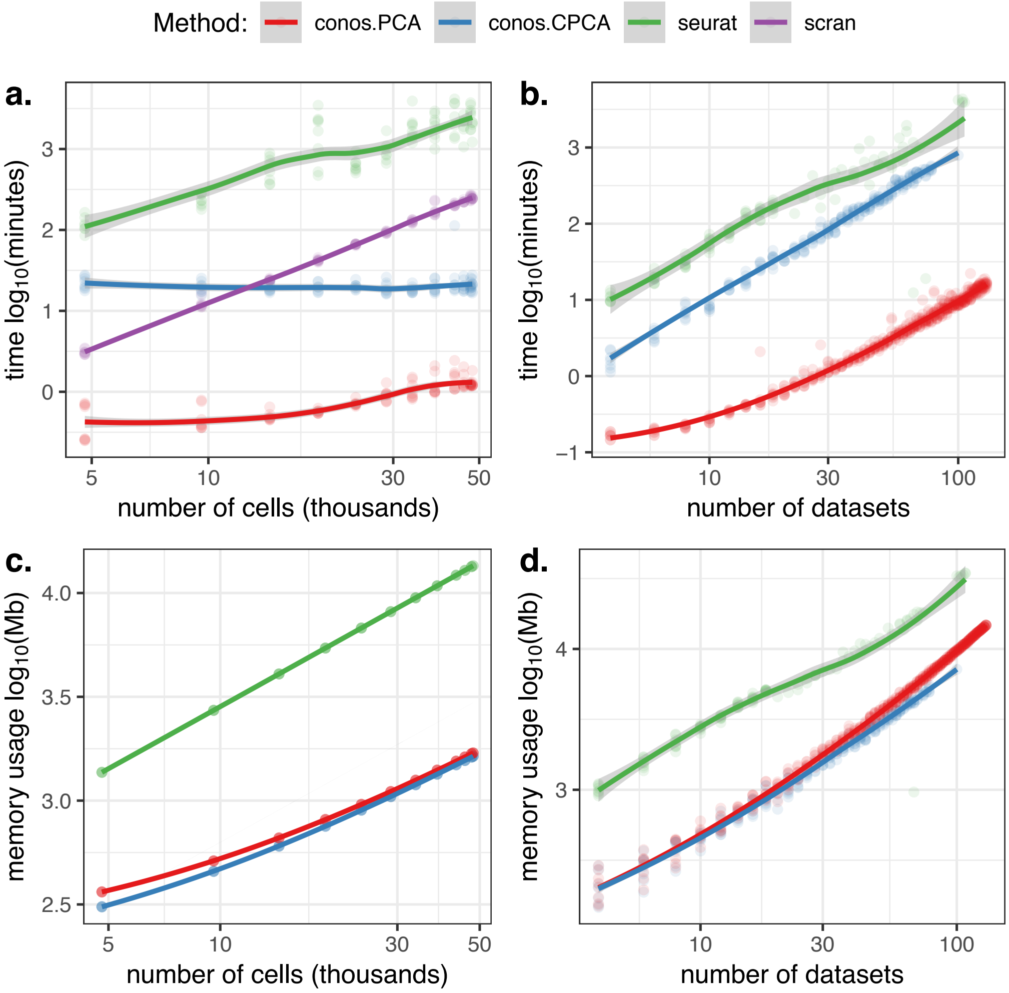
Supplementary Figure 16. Agreement of *Conos* clustering with the published Tabula Muris annotations. The agreement is illustrated using dot plots, with the size of the dot corresponding to the number of cells intersecting between a given cluster and a published annotation category, and the color specifying Jaccard coefficient.



Supplementary Figure 17. Agreement of *Conos* clustering with the published Han *et al.* annotations. The agreement is illustrated using dot plots, with the size of the dot corresponding to the number of cells intersecting between a given cluster and a published annotation category, and the color specifying Jaccard coefficient.



Supplementary Figure 18. Published annotation of two mouse atlases on an embedding of a joint graph optimized using decoy cell alignment. Similar to the Supp. Fig. 15, the panels show annotation of the joint graph combining Tabula Muris and Han *et al.* mouse atlases (173 datasets >400k cells), using the annotation from the original publications of Tabula Muris (**a.**) and Han *et al.* (**b.**) The top left inset shows overview of complete joint embedding (all 174 datasets), and right inset shows distribution of the platforms within the joint embedding (green: Smart-seq2, blue: microwell, red: 10x Chromium). The bottom right insert shows the schematic logic of the decoy cell refinement: based on the initial low-resolution clustering of a joint graph, Conos identifies major subpopulations that are missing from different samples (in the case, Sample B was missing a blue subpopulation). Conos then reruns pairwise sample comparisons, adding blue decoy cells into Sample B when comparing it with samples that have blue subpopulations in them (like Sample A). The edges mapping to the decoy cells are then discarded, giving cleaner graph.



Supplementary Figure 19. Runtime and memory performance. a. The plots show average runtimes for the cell subsampling benchmark (16 HCA BM+CB datasets were downsampled from 3k cells each to lower numbers of cells; n=10 downsampling rounds were performed for each point). The CPCA performance stays constant, as the runtime complexity of the CPCA fit depends critically on the number of samples, not on the number of cells. **b.** Runtime complexity of combining increasing number of datasets, each containing 1k cells. **c,d.** Memory usage under the scenarios shown in panels a. and b. Note: the benchmarking scripts did not estimate embeddings, which would lead to additional CPU and memory load. The shading shows the 95% confidence region of the mean.

Supplementary Note 1. Clustering Stability Under Perturbations

Systematic difference in the output of clustering algorithms

Clustering serves as a convenient grouping of cells, facilitating further interpretation and navigation of the datasets. It is important to point out that just like in most other contexts, clustering of cells is an approximation, and does not have a unique solution. Consider clustering of the HCA BM+CB dataset using two different community detection (graph clustering) methods (Figure 1).

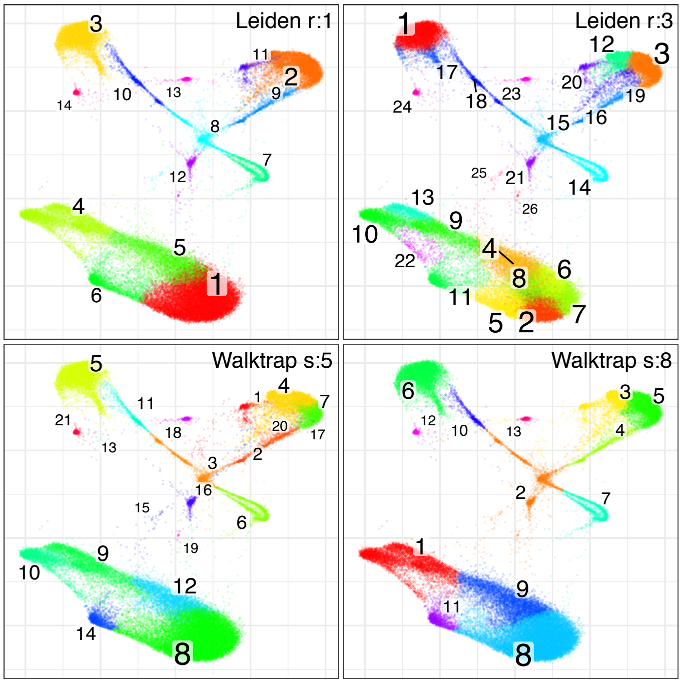


Figure 1. Variation of clustering granularity. The four panels show clustering of the integrated HCA BM+CB dataset using two different clustering methods (Leiden and walktrap community detection algorithms) each with two different parameter settings (varying "resolution" parameter for Leiden clustering, "steps" for walktrap). The resulting clusters can split and merge different subpopulations and systematically vary in their characteristic size. Algorithmically, there is no clear single "correct" resolution.

Both algorithms are designed around heuristic optimization of certain network features. While one algorithm returns more fine-grained clusters than others, it is usually impossible to claim that one of them is better without bringing in additional biological knowledge.

The relationships between the cells can be usually approximated better using hierarchical clustering. For instance, one may separate a cluster of T lymphocyte, and then within it CD8+ naïve or cytotoxic T cells. Or CD4$^+$ Treg and Th subpopulations, *etc*. The coarse clusters (such as a cluster combining all T lymphocytes) represents a real and, likely, stable grouping of cells. However, as one considers more fine-grained groupings (small clusters), at some point the ability to resolve different subpopulations will approach the noise of the observations, and the clusters will become unstable. Here we discuss various measures for assessing cluster stability under a simple subsampling perturbation (*i.e*. rerunning clustering on just 95% of cells).

Stability of simple partition clustering.

We first consider the simplest case, when clustering returns only a flat partition of cells into a set of non-overlapping clusters. Evaluation of agreement between the original clustering result and a result on a subsampled dataset (95% of cells) can be performed using standard measures, such as the adjusted Rand index (aRI). One can also assess cluster-specific stability measures, such as Jaccard coefficient of the optimally matching cluster (JC)[Hennig, 2007 #90]. The plot below shows both cluster-specific JC and combined aRI for the HCA BM+CB dataset, for the partition determinized by the Leiden clustering algorithm, based on 100 such randomized subsampling runs.

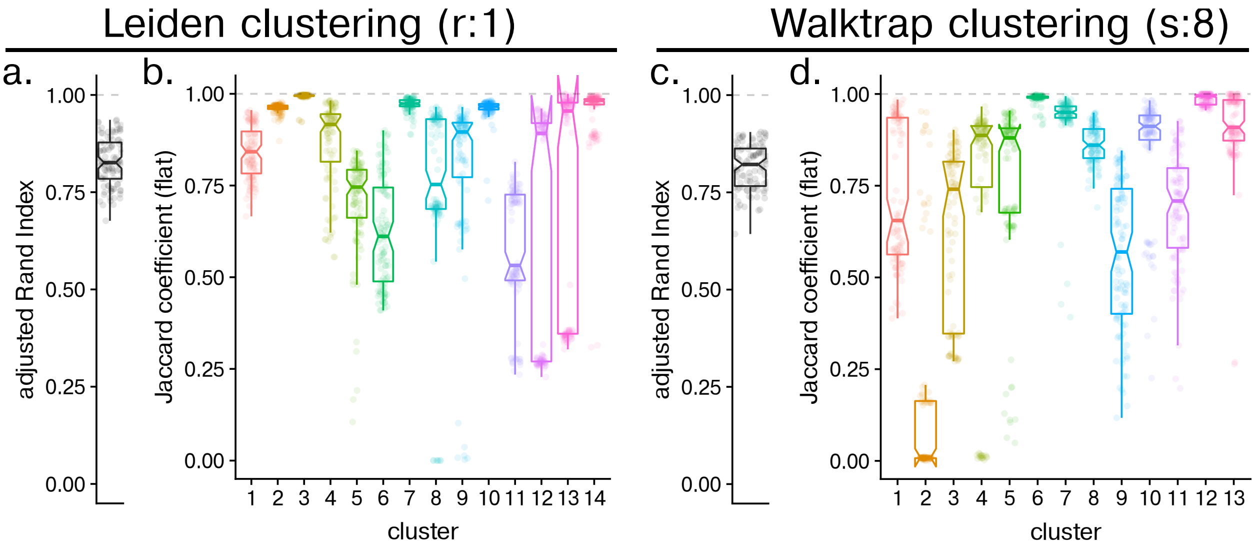


Figure 2. Stability measures on flat partition clustering. Given a simple flat partition of cells into clusters, the stability is assessed by repeating the clustering procedure on the subset of cells (95% of cells) and calculating standard agreement coefficients: **a.** adjusted Rand index, **b.** Jaccard coefficient, calculated for each cluster relative to the best-matching cluster in the subsampled partition. Results of 100 randomized subsampling rounds are visualized using boxplots. Panels a,b show the results for the Leiden community detection algorithm, and panels **c,d** for the walktrap community detection method. In this plot and all others, the boxplot center shows median, upper/lower box lines mark top 75% (Q3) and bottom 25% (Q1) levels; Whiskers extend from max(min(x),Q1-1.5 IRQ) to min(max(x),Q3+1.5 IQR, where IRQ is the inter-quartile range. The notch shows 95% confidence interval of the median.

The example above shows good overall stability (aRI), however cluster-specific stability of cluster 2 in panel d. (and some others) are low. This can be due to several reasons:

1. Cluster 2 is a noise artifact and groups random cells that don't have a tendency to form a separate cluster.
2. Cluster 2 represent a smaller grouping of cells that's not typically distinguished at this level of resolution, but may a stable feature if the resolution of clustering is increased.

To distinguish between the two, one must scan through a range of resolutions. A good way of doing this is to consider hierarchical clustering methods.

Stability of a partitioning cut in hierarchical clustering.

Hierarchical community detection algorithms, like walktrap, report the entire hierarchy, where the leaves of the dendrogram correspond to the individual cells. The optimal cut is then determined based on some criteria (modularity optimization, in the case of walktrap), resulting in a flat partition assignment of cells. Given the set of clusters comprising the flat partition, one can use various measures to evaluate to what extent any given cluster is matched by some subtree in the dendrogram derived from a perturbed (95% subsampled) dataset. The boxplots below (Figure 3) show statistics for the walktrap clustering algorithm. The cluster stability evaluation is performed comparing the original flat partition to the flat partitions in each perturbation (left), or comparing the original flat partition to the full dendrogram of each perturbation (right):

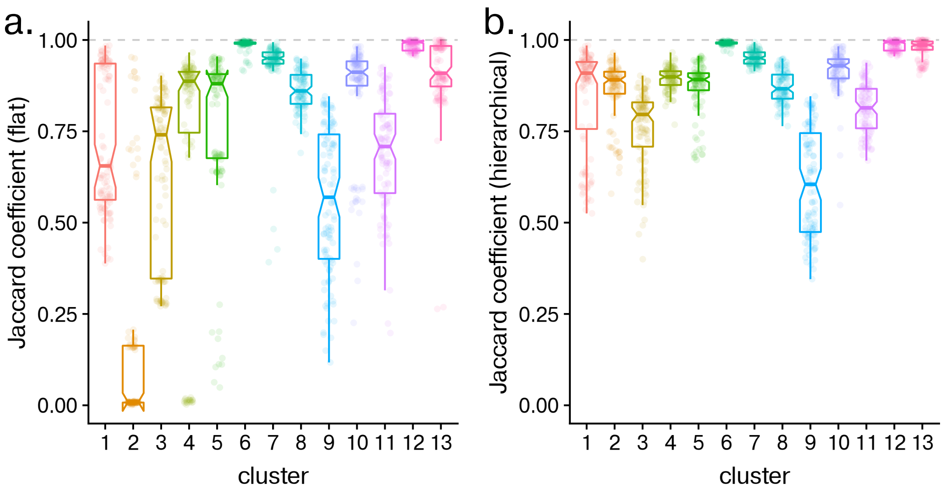


Figure 3. Stability based on hierarchical clustering. **a.** Distribution of Jaccard coefficients calculated relative to flat partition results of the walktrap method (same as in Fig. 2b). **b.** Distribution of Jaccard coefficients for the same clustering, but calculated relative to an optimal subtree within the hierarchical result of the walktrap algorithm.

As expected, when entire tree is considered, the results improve, as the benchmark is no longer affected by unstable cluster split/merge effects as these differences simply represent different levels in the hierarchical clustering.

Hierarchical view of cluster stability.

The hierarchical considerations are also useful to consider with respect to the reference clusters (that are being tested for stability). These can also be considered in a hierarchical context, and various stability measures can be evaluated not just for the leaves (*i.e.* individual clusters), but internal nodes representing combinations of related clusters.

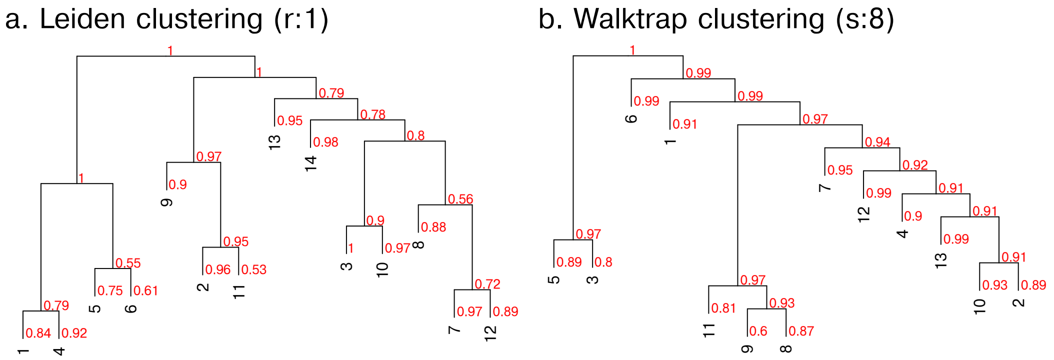


Figure 4. Jaccard coefficients for dendrograms of clusters. **a.** Hierarchical relationships among the 14 clusters determined by Leiden (resolution=1) algorithm is shown. The median of optimal Jaccard coefficients (calculated relative to similar hierarchies reconstructed from 95% subsampling rounds) is shown next to every leaf and internal node in red. **b.** Analogous view of the walktrap (steps=8) results.

This view can illustrate that while some of the lower-level nodes may not be very stable, their combinations can be very stable. For example, for the Leiden (r=1) results (see Figure 1), the clusters 1,4,5,6 represent different subpopulations of T cells. While stability of the individual clusters (relative to Leiden clustering on the subsampled dataset) can be marginal, their combination is always detected (Figure 4a).

The cluster dendrograms can be derived in a variety of ways. The analysis above (Figure 4) used upper part of the hierarchical clustering reported by walktrap to determine the dendrogram for the walktrap clusters. To determine a hierarchy for the Leiden clustering, the joint graph was simplified by collapsing all the nodes (cells) belonging to the same cluster and combining corresponding edges. The hierarchical clustering of the resulting small graph was then calculated using walktrap algorithm.

Stability under parameter perturbations

In constructing the joint graph, on which the clusters (node communities) are determined, *Conos* employs a number of parameters. Here we use example of the BM+CB 16-dataset collection to analyze sensitivity to these parameters.

Neighborhood size *k*. The neighborhood size parameter *k* is used to determine the size of the neighborhood considered during inter-sample comparisons. As such, it directly influences the number of the resulting inter-sample edges. Under the default mutual-nearest neighbor matching, relationship depends on other factors, such as subpopulation sizes, homogeneity, and the magnitude of the batch effect. There is no obvious optimal value of *k*. One can select *k* to optimize the overall modularity of the resulting clustering (see Figure 5), however it is unclear to what extend this heuristic would reflect the biologically meaningful integration of samples (*e.g.* a situation where all the samples remain well-separated may end up having higher modularity).

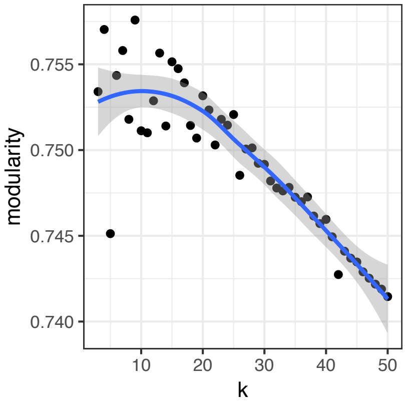


Figure 5. Dependency of the total modularity of the multilevel partition on the neighborhood size k. Multilevel community detection was ran on the HCA BM+CB 16-dataset collection, calculating the overall modularity of the resulting multilevel.community partition using igraph::modularity(). Smoothed estimate of the mean is shown with shading indicating the 95% confidence interval.

Larger values if *k* connect samples more densely, however we expect major community structure to remain the stable. To illustrate that, we have used HCA BM+CB example to rerun graph construction using different values of *k*, comparing the similarity of the resulting clustering (using flat partition-based adjusted Rand index, or mean Jaccard coefficient in a hierarchical comparison) between different values of *k*. The results (Figure 6) show that clustering stabilizes once a minimal value of *k* reached to establish reasonable connectivity of the graph. Importantly, further increases in *k* do not disrupt the structure of the major communities in the graph. Specifically, the stability of the resulting clustering is comparable to the base-level stability of the clusters, as assessed from 95% subsampling of the cells (Figure 6).

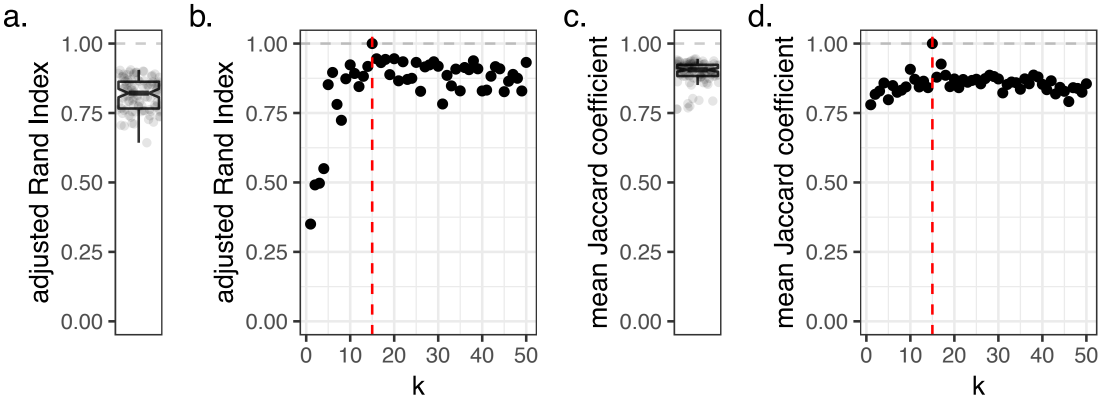


Figure 6. Clustering stability with respect to variations in parameter k. a. As a reference, stability of the clustering with respect to a 95% random subsampling of the cells is shown using adjusted Rand Index (same as in the left panel of Figure 2c). b. Adjusted Rand index as a function of *k*. The vertical red line gives the position of the reference point (*k*=15). Clustering stabilizes beyond minimal values of *k.* c. As a reference, stability of the clustering with respect to a 95% random subsampling of cells is quantified as a mean Jaccard coefficient across clusters, calculated based on the best matching subtree (as in the right panel of Figure 2c, averaged across clusters). d. Mean Jaccard coefficient as a function of *k*. All clustering analysis were performed using walktrap.community() method with steps=8.

Within-sample neighborhood size *k.self*. Analogous to the *k* parameter in the inter-sample comparisons, *k.self* determines the number of neighbors used to establish within-sample edges. These edges are mostly meant to tie in cells that end up with no or very few inter-sample edges into the joint graph, and should be kept at minimal levels (not to increase influence of dataset-specific subpopulations). Here as well, the results show that the stability with respect to variations in *k.self* parameter remains at the base-level, as determined by the 95% subsampling of cells.

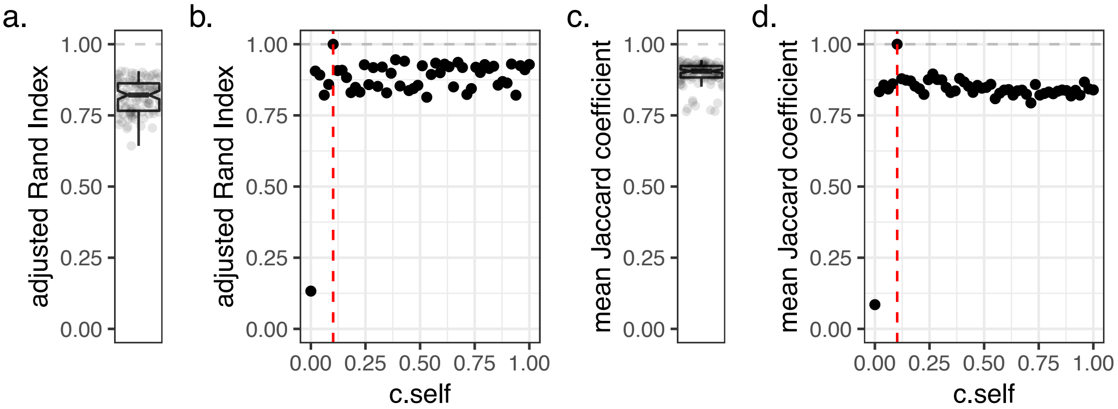


Figure 7. Stability with respect to k.self parameter. Analogous to the Figure 6, the plots compare the stability of the clusters due to variations in *k.self* parameter (b,d) with the base-level stability of the clusters as determined by the 95% subsampling of cells (a,c).

Relative weight of the within-sample edges (k.self.weight). To ensure that the major structure of the graph is driven by inter-sample relationships, *Conos* by default maintains lower weight of the within-sample edges using k.self.weight=0.1 (note, k.self.weight parameter is referred to as

c_{self} in the Methods section). The plots below show sensitivity to the exact value of this relative weight parameter.

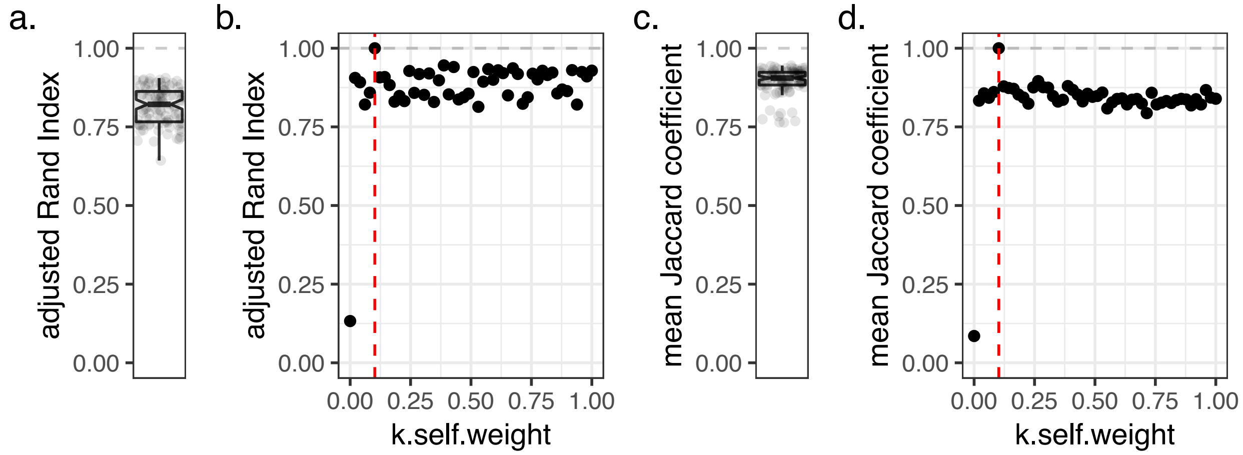


Figure 8. Stability with respect to k.self.weight parameter. Analogous to the Figure 6, the plots compare the stability of the clusters due to variations in *k.self.weight* parameter (b,d) with the base-level stability of the clusters as determined by the 95% subsampling of cells (a,c). The red dashed line marks the default value of 0.1.

Number of principal components. The number of principal components (PCs or CPCs) can influence the results. Clearly, selecting too few components will degrade the ability to identify subpopulations as a substantial amount of variance would be left unexplained (Figure 8). Selecting too many components should not be a problem, particularly since by default *Conos* uses correlation-based distance measure which is robust to increasing number of components.

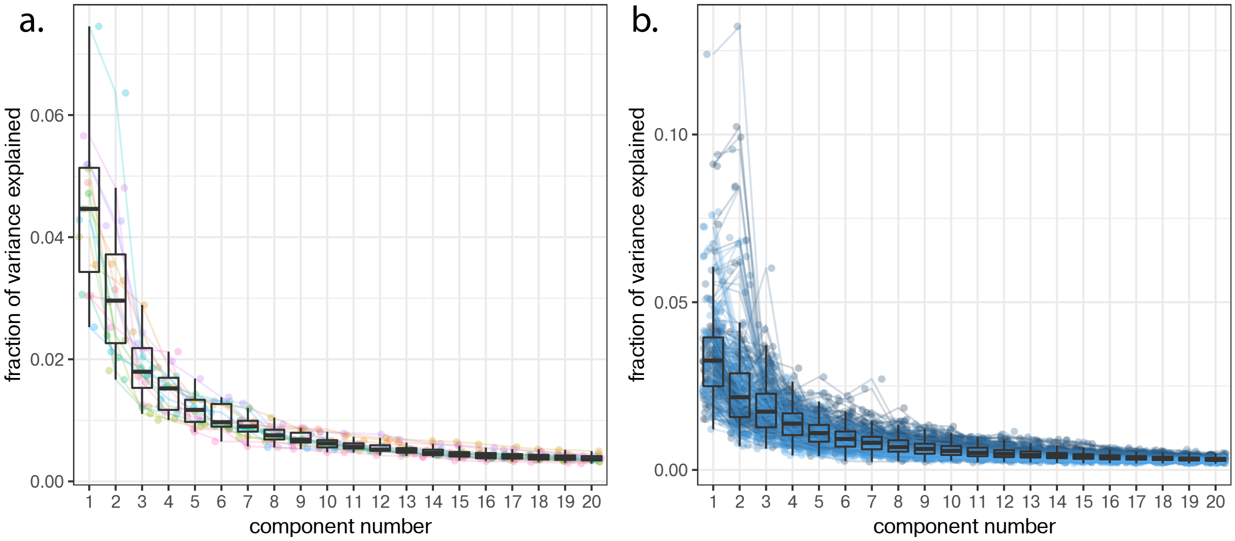


Figure 9. Fraction of variance explained by principal components. **a.** Fraction of variance explained by top principal components (PCs). Colored lines show values for each dataset, with boxplots summarizing the values across datasets. **b.** Fraction of variance explained by top common principal components (CPCs). Colored lines show variance explained by CPCs in each pairwise dataset comparison, with the boxplots showing summaries across all pairwise comparisons. HCA CB+MB dataset was used. The plots were constructed using conos::plotComponentVariance() function. Boxplots show

Evaluating the stability of the clustering with respect to the number of utilized components (Figure 9), shows the expected picture, with the results stabilizing once some minimal number (e.g. 12) of components is reached.

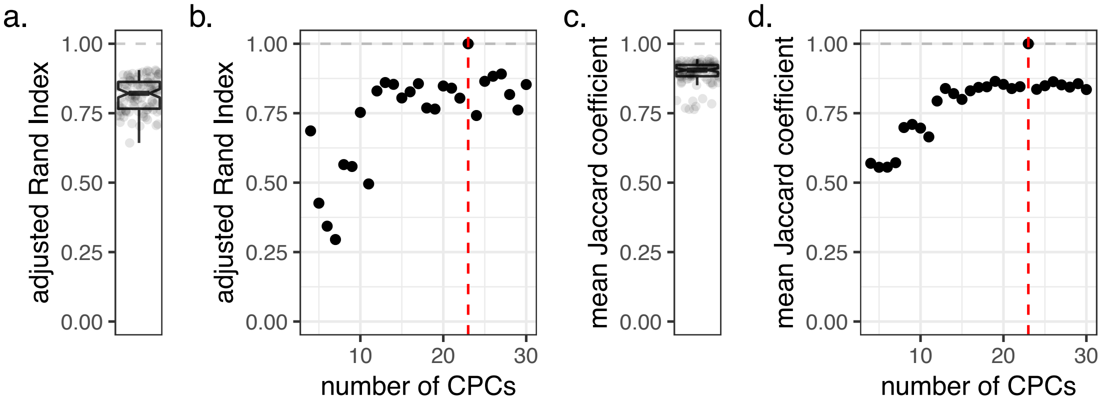


Figure 10. Stability of the clustering with respect to the variation in the number of principal components. Using the same layout as in Figure 6, the plots compare the stability of the clusters due to variations in the number of top CPCs used (b,d) with the base-level stability of the clusters as determined by the 95% subsampling of cells (a,c). As expected, the clustering stabilizes after some minimal number of components are taken into account.

Supplementary Note 2. Integrating RNA-seq and ATAC-seq datasets

Integration between distinct modalities, such as transcriptional and epigenetic measurements is a challenging topic that introduces additional technical considerations. Here we apply *Conos* approach to integrate an ATAC-seq based panel of measurements, and then show integration between ATAC-seq and RNA-seq datasets. We illustrate that although *Conos* such integration can be quite effective, its success depends on the resolution of the data and ability to find an informative link between gene expression and other modalities.

Integration of multiple ATAC-seq datasets

To introduce chromatin accessibility data, we first show integration of 17 sci-ATAC-seq replicates covering 13 mouse tissues[1]. We integrate the data based on accessibility-based gene activity scores[2], feeding them into *Conos* in the same manner as is normally done for RNA-seq. The resulting integration clearly separates distinct cell types, joining analogous cell types across tissues and replicates where appropriate (Figure 1).

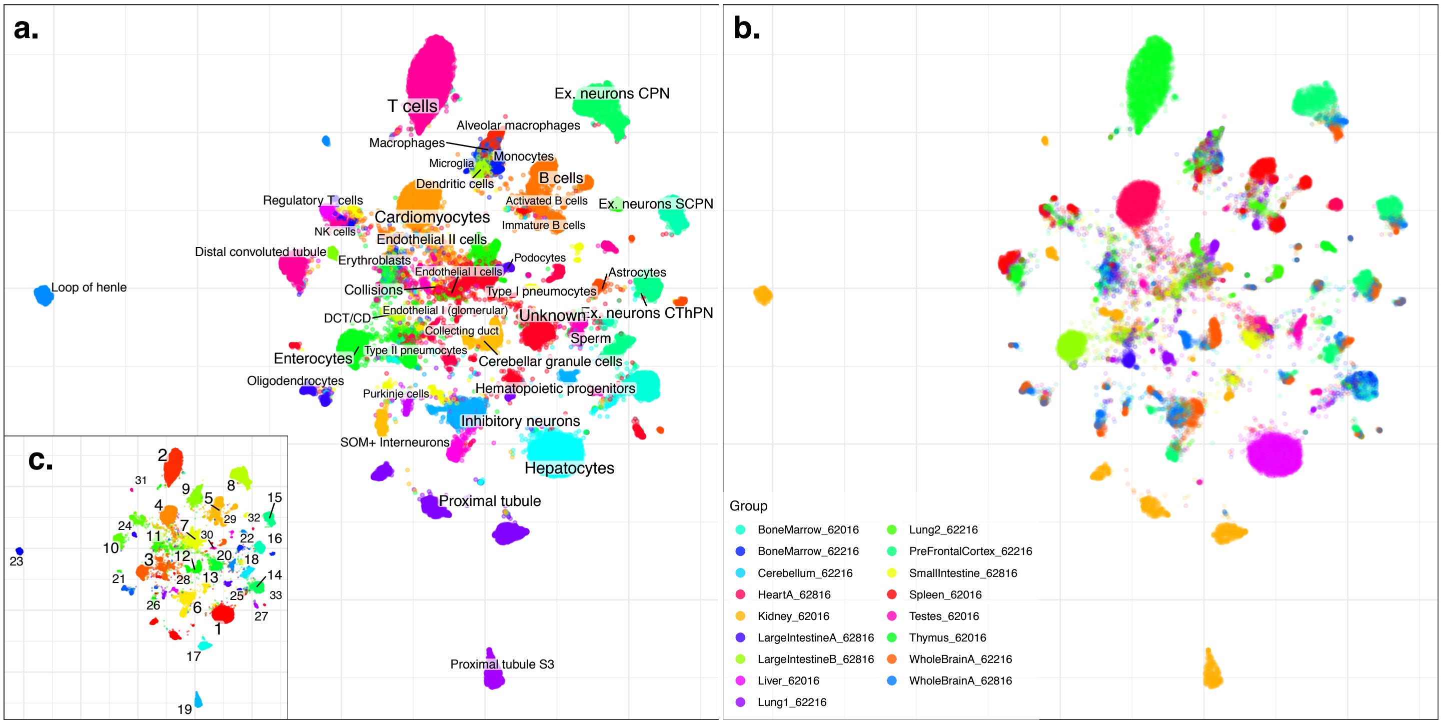


Figure 1. Integration of multiple sci-ATAC-seq datasets. **a.** largeVis embedding of the joint graph integrating data from 17 sci-ATAC-seq replicates are shown, colored and labeled according to the annotation provided in the original publication[1]. **b.** Coloring by batches (see legend). Most of the batches sample distinct tissues and do not intersect. The two bone marrow replicates show expected mixing. The whole brain replicates are appropriately grouped by cell types (typically together with the prefrontal cortex sample), however retain some batch specificity within each cell type. **c.** *Conos* clustering of the joint graph separates expected cell types and integrates across replicate batches.

Integration of scRNA-seq and sci-ATAC-seq

To demonstrate integration between different modalities we focused on specific tissues that were covered in the mouse transcriptional atlases[3,4]. First, we consider lung, using *Conos* to integrate the two sci-ATAC-seq replicates for that tissue, together with three scRNA-seq datasets produced using different platforms (Smart-seq2, 10x Chromium, and microwell technique).

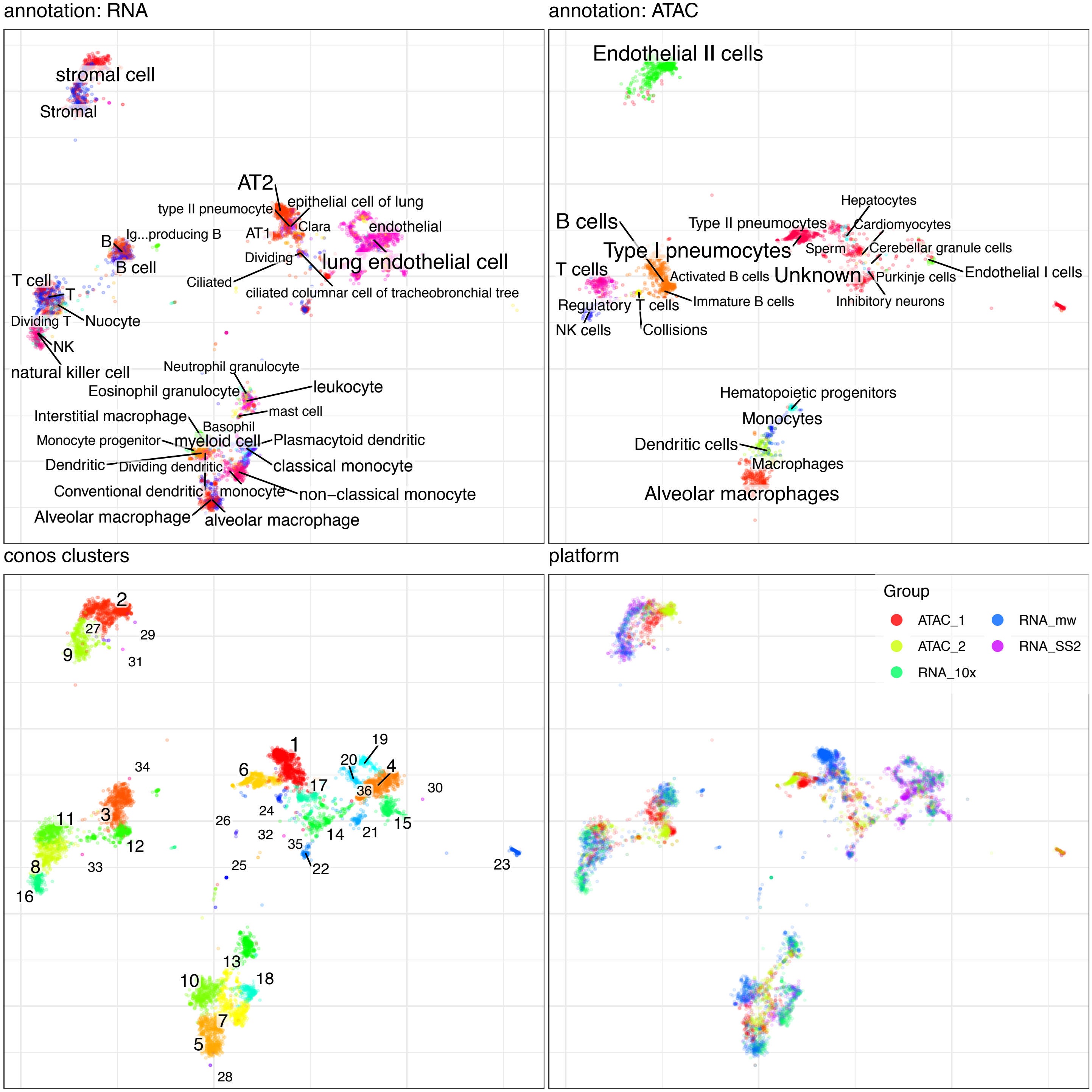


Figure 2. Integration of scRNA-seq and sci-ATAC-seq data on the lung. **(top left)** scRNA-seq annotations (as provided in the original publications) are shown on the joint embedding. Only RNA-seq cells are shown. **(top right)** sci-ATAC-seq annotations, as provided by the original publication are shown on the joint embedding. Only sci-ATAC-seq cells are shown. **(bottom left)** Cell clusters, as determined by *Conos* on the joint embedding. **(bottom right)** Joint embedding, colored by the dataset identity of each cell.

Conos effectively integrates the RNA and accessibility data on the lung (Figure 2), based on the gene-level accessibility summary scores derived using *Cissero*[2], aggregating major cell types (e.g. macrophages, T, B, endothelial cells), as well as subtle differences such as that between T and NK cells. Note that we show all annotations provided for all integrated cells in the original publications, and some of them appear to include erroneous labels for small groups of cells (*e.g.* Sperm, Cardiomyocytes) stemming from the aggregation method used. We next performed similar analysis for bone marrow datasets (Figure 3).

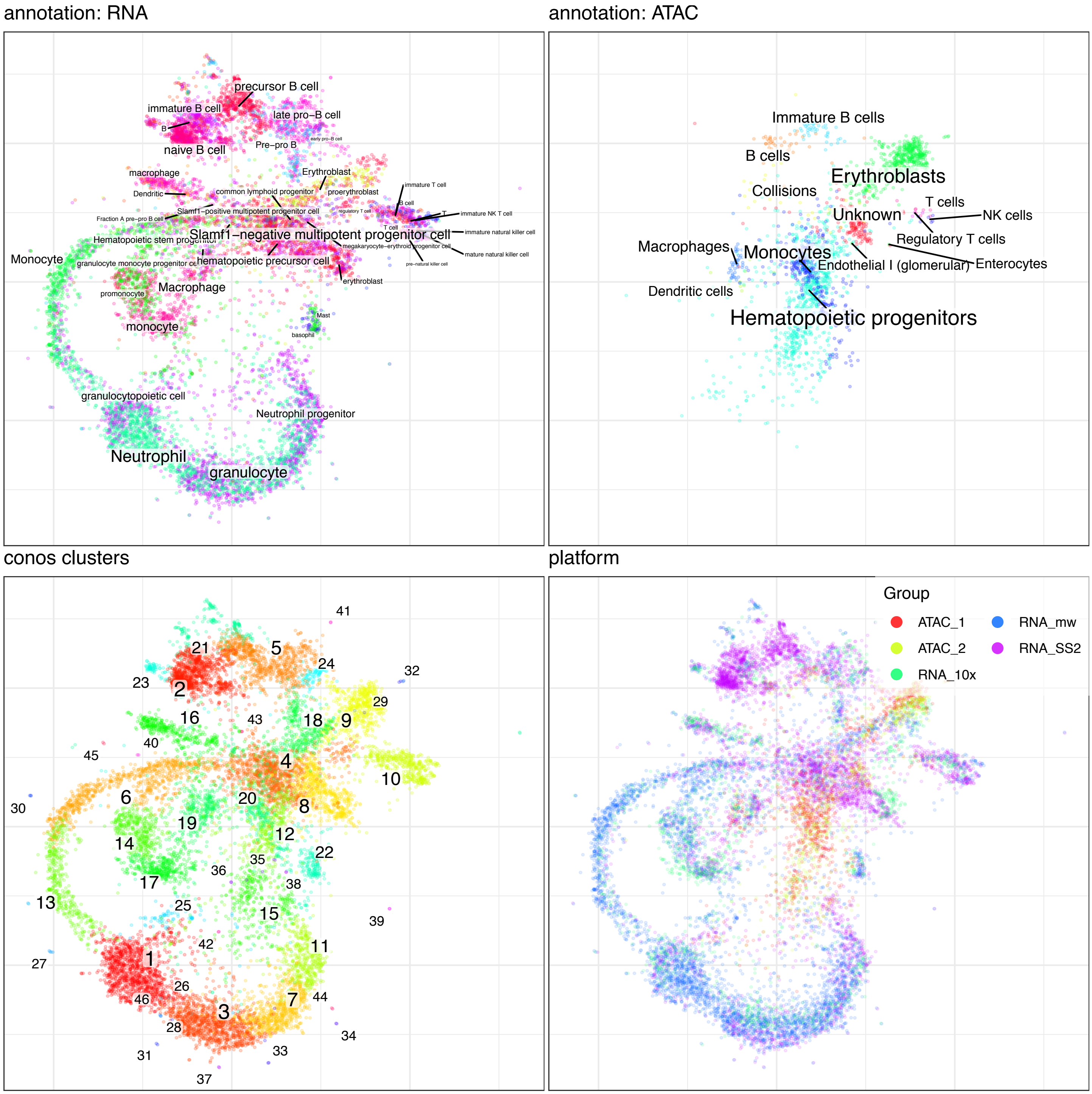


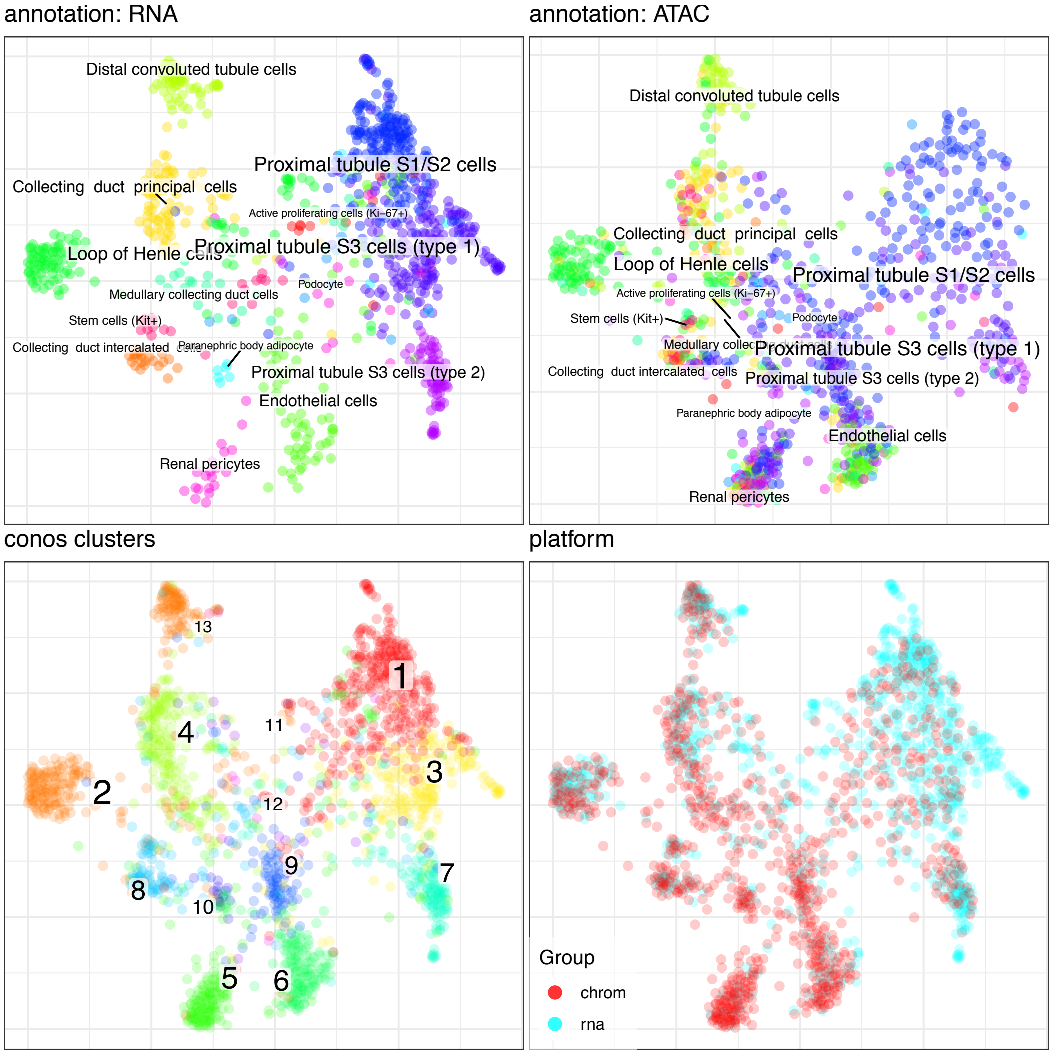
Figure 3. **scRNA-seq and sci-ATAC-seq integration of the bone marrow data.** Similar to Figure 2, the panels show original RNA-based and ATAC-seq based annotations (top panels), *Conos* clusters, and platform factors on the joint embedding.

The cell type representation of the bone marrow differs notably between the sci-ATAC-seq and scRNA-seq, with scRNA-seq data showing extensive granulocyte maturation trajectory, and sci-ATAC-seq data focused on the hematopoietic progenitor population. Nevertheless, *Conos* correctly aligns the overlapping cell types, including erythroblasts, T, B cells, macrophages, and hematopoietic progenitors. It also resolves the difference between immature and mature B cell clusters. The mapping of the progenitor populations is slightly shifted, however that could also be due to the representational biases.

Evaluating RNA-seq/ATAC-seq alignment consistency and limitations using sci-CAR data

While the results above illustrate the general ability to align across modalities, to get a more precise idea about the mapping correspondence we examined data from the sci-CAR technique that measures both transcriptomes and chromatin accessibility in the same cell. In this case, the true alignment across the datasets is in essence established by the barcode identity of the RNA-seq and ATAC-seq data, making for a convenient benchmark.

The joint nature of the sci-CAR technique represents an impressive technical feat, but the molecular coverage achieved for each cell is lower, which is particularly notable in the analysis of the chromatin accessibility aspect. As was shown in the original publication, however, more robust accessibility profiles can be constructed by aggregating molecules across cells with similar transcriptional profiles. For the first analysis below, we used RNA-seq based clustering of the cells to partition ATAC-seq cells into groups of 10 cells based on the similarity of their transcriptomes, and then combined all the data within each group of 10 cells to obtain “meta-cells” with 10x coverage (note, the original publication used 50x aggregation). We then summarized the chromatin accessibility at a gene level as a sum of all accessibility signal at the detected peaks (as defined in the original publication) across the entire gene body and the 10kb margins around the gene. Feeding such matrices into *Conos*, and increasing *k* to 200 to get a more focused mapping results in a reasonable alignment of the two modalities (Figure 4).


annotation: RNA

annotation: ATAC

conos clusters

platform

Figure 4. **Integration of RNA and accessibility data from sci-CAR mouse kidney measurement.** Using the panels analogous to those shown in Figures 2 & 3, the plots show results of *Conos* integration between RNA and accessibility modalities measured using sci-CAR, using 10x aggregation of the ATAC-seq cells.

On a population level, *Conos* integration performs well, aggregating RNA- and accessibility- based measurements on Renal pericytes, Loop of Henle cells, Endothelial, Collecting duct, and Distal convoluted cells. The abundant Proximal tubule S1/S2 and S3 subtypes also show corresponding alignments, though quite a few of these cells are mismapped to other cell types. The obvious advantage of using sci-CAR is that we can explicitly quantify this performance. Here we use the normalized rank of the true corresponding cell based on the Euclidean distance in the resulting embedding as the measure of performance (Figure 5).

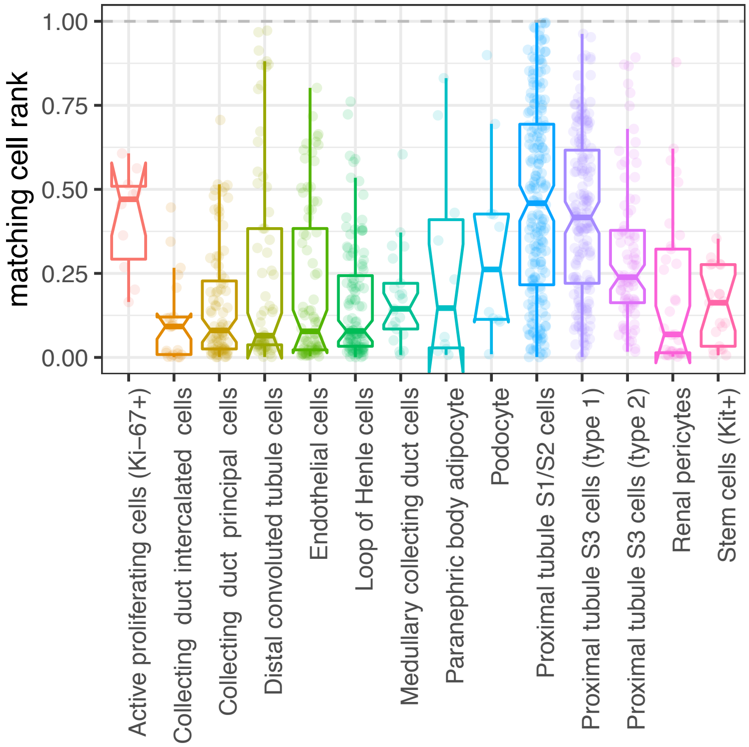


Figure 5. Performance of RNA-ATAC-seq alignment based on whole-gene summary scores. The boxplots show, for each cell type, the distribution of the rank of the corresponding RNA cell in terms of Euclidean distance in the embedding from the ATAC-seq version of the same cell. An ideal alignment would have all ranks at 0. A random alignment would be closer to ½. The performance varies by cell type, with the proximal tubule and proliferating cells showing worst performance. In this plot and all others, the boxplot center shows median, upper/lower box lines mark top 75% (Q3) and bottom 25% (Q1) levels; Whiskers extend from max(min(x),Q1-1.5 IRQ) to min(max(x),Q3+1.5 IQR, where IRQ is the inter-quartile range. The notch shows 95% confidence interval of the median.

The reason why *Conos* alignment works is because the devised per-gene accessibility summary scores show some linear correlation with the gene expression. Therefore, we expect the choice of the scoring scheme to be critical. To illustrate that, we will use another (reasonable) summary measure on the same dataset: a total accessibility in 10kb or 4kb region around TSS of each gene (Figure 6). As it can be seen, the performance of the TSS measure is worse than that of the whole-gene measure, with 4kb measure showing near-random performance

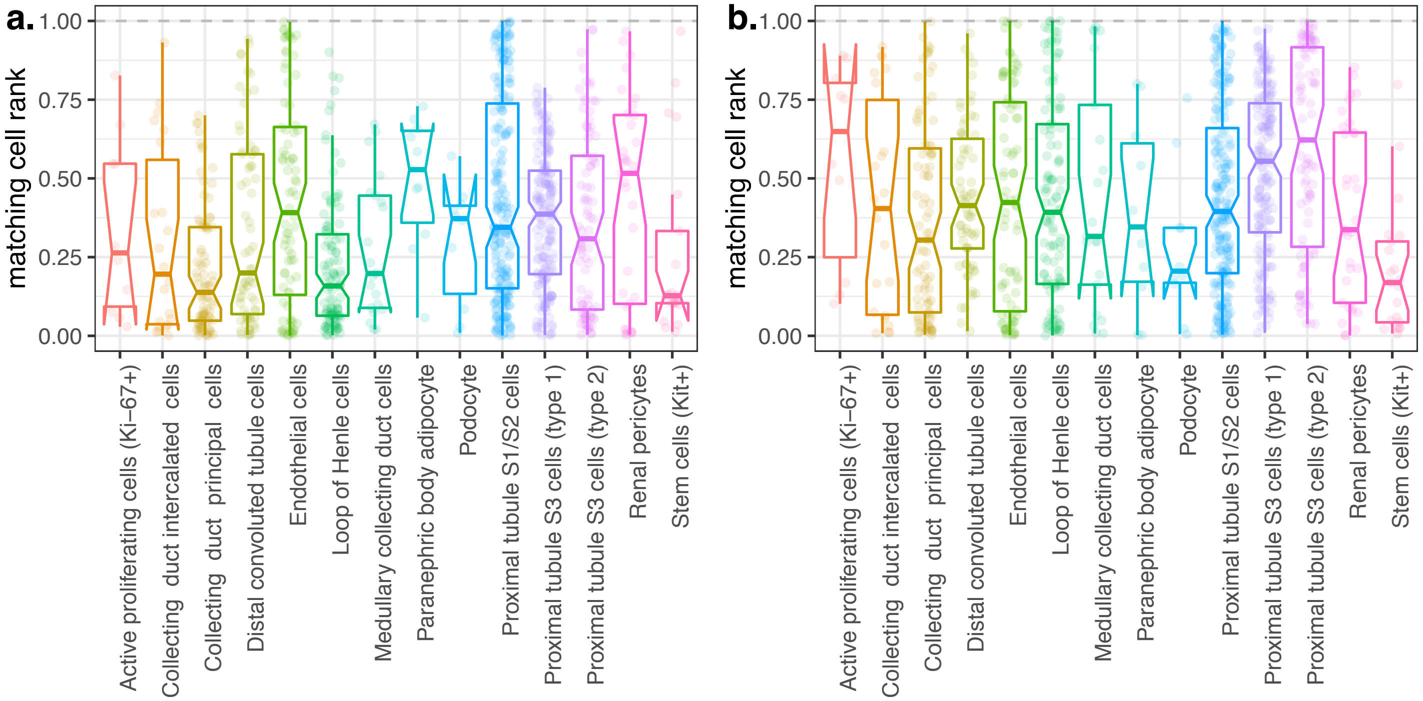


Figure 6. Alignment performance using gene-level accessibility summaries with windows of different size. To illustrate how alignment depends on the method for summarizing gene-level accessibility, we the plots show matching cell ranks (same as Figure 5) for alignments where ATAC-seq data was summarized as total signal within (**a.**) 10kb around TSS, or (**b.**) 4kb around TSS. The later, clearly provides too little signal and the cells are not well aligned.

References

1 Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C. & Shendure, J. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e1318, (2018).

2 Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J. & Trapnell, C. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858, (2018).

3 Tabula Muris, C., Overall, c., Logistical, c., Organ, c., processing, Library, p., sequencing, Computational data, a., Cell type, a., Writing, g., Supplemental text writing, g. & Principal, i. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372, (2018).

4 Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G. C., Chen, M. & Guo, G. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107 e1017, (2018).