**3.4 Annotating clusters**

As mentioned at the start of this chapter, cell clustering has allowed biologists to define cell types in an unsupervised manner. Thus, an important end goal of cell clustering is to annotate the identified cell clusters. Cell type annotation / classification is not a trivial task and often require expert opinion in the tissue being studied. However, computational tools have been developed to leverage on large volumes of existing single-cell data to help accelerate the process of cell type annotation.

**3.4.1 Different approaches to cell type annotation**

There exist many computational tools that can assign cell type labels to single cells as described by this benchmarking study ([Abdelaal et al. 2019](https://ouyanglab.com/singlecell/clust.html#ref-abdelaal2019_cellID)) and this review ([Pasquini et al. 2021](https://ouyanglab.com/singlecell/clust.html#ref-pasquini2021_celltypeReview)). Briefly, there are three main computational approaches to annotate cell types (Figure @ref{clust-celltype}). First, marker gene approaches leverage on lists of marker gene sets that can distinguish different cell types from existing literature and previous single-cell studies. Single cells or clusters are then scored using these marker gene sets e.g. using AddModuleScore to estimate the overall expression levels of these genes. Some heuristics / scoring criteria is then applied to assign the cell type. Second, correlation approaches apply multiple correlation measures to estimate the similarity between the single cells / clusters in the input dataset against some reference data e.g. single-cell atlases or bulk RNA-seq databases such as GTeX / FANTOM. Third, supervised classification approaches uses machine learning or deep learning to predict cell type labels. The learning model is first trained on some single-cell reference atlas before being applied onto the input dataset to compute the probability that a single cell / cluster belong to a particular cell type.

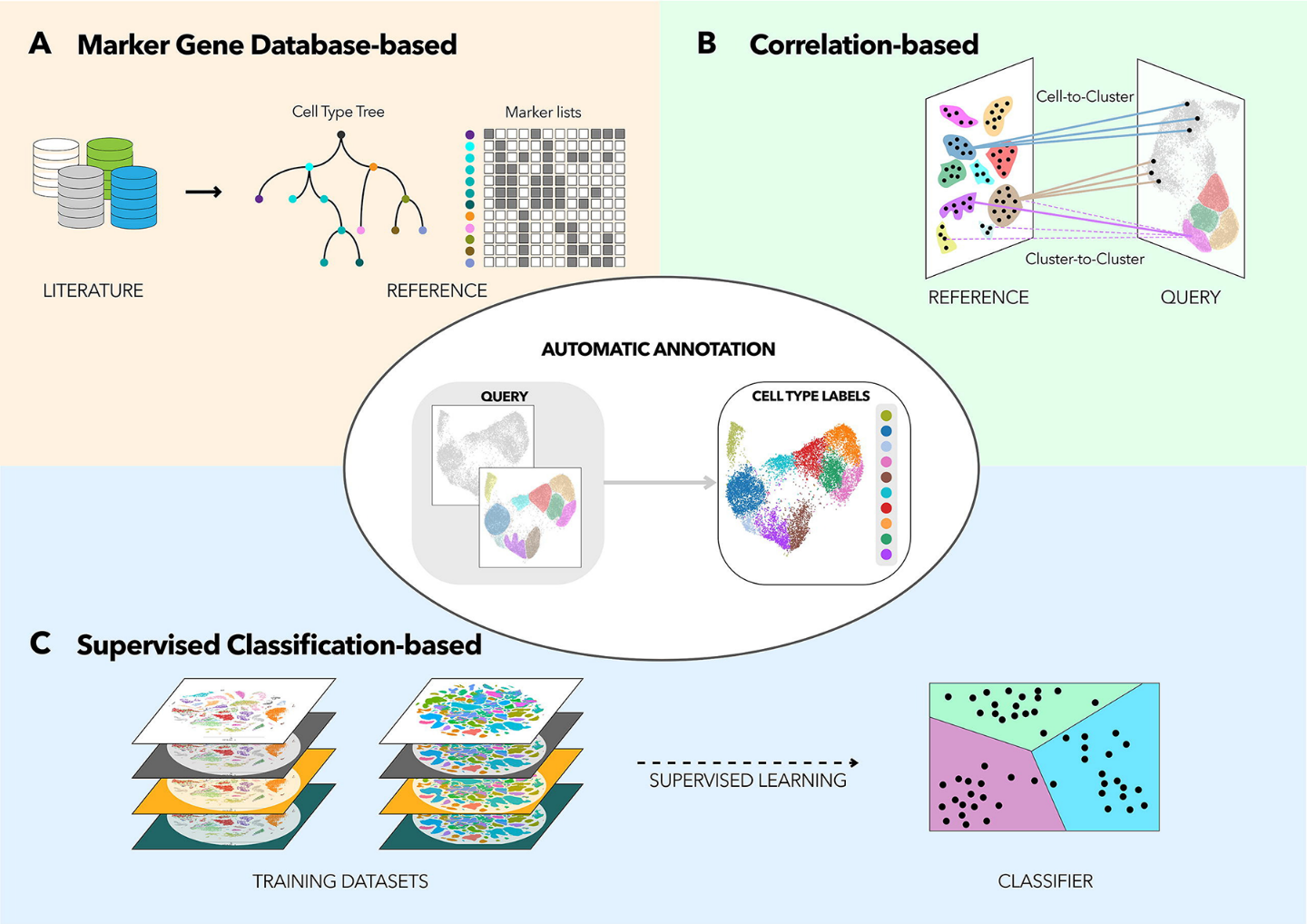


Figure 3.14: There are three main approaches to cell type annotation in single-cell studies, namely marker gene approaches, correlation approaches and supervised classification approaches. Image taken from Pasquini et al. ([2021](https://ouyanglab.com/singlecell/clust.html#ref-pasquini2021_celltypeReview))

In all three approaches, the choice of the reference data is crucial for the success of the cell type annotation. Ideally, the reference data should be encompass the tissue / cell types that are profiled in the input dataset. Otherwise, spurious results may occur e.g. when one tries to annotate the bone marrow dataset using a brain atlas reference. Furthermore, the cell type labels in the reference data will determine the “resolution” of the predicted cell type annotation. For example, using a reference that only has coarse cell type labels (e.g. having a general T-cell label as opposed to CD4+ Naive, CD8+ Naive etc finer labels) can only result in coarse-level annotations and vice versa.

Also, it is important to have the option of having unassigned cells. This happens when a single cell / cluster has a very low score across all the cell types present in the reference. Biologically, this can be a novel rare cell population that is not annotated in the reference data or represent stressed / dying cells that has “lost” their cell fate identity. Furthermore, cell fate is highly plastic and there is a possibility that cells are trans-differentiating between two cell fates. In this scenario, it is possible that a single cell / cluster has high scores in two different cell types.

**3.4.2 Annotation via marker genes and functional analysis**

In this guide, we will be annotating the clusters manually based on the marker genes and HAY\_BONE\_MARROW gene signature overrepresentation analysis that we have performed earlier (Figure @ref{clust-annotTable}). The major cell state e.g. HSC / progenitors / erythroid / dendritic cell / monocyte / B cell / T cell / NK cell can be determined from the HAY\_BONE\_MARROW signatures and some of the cellular substates can be resolved using marker genes.

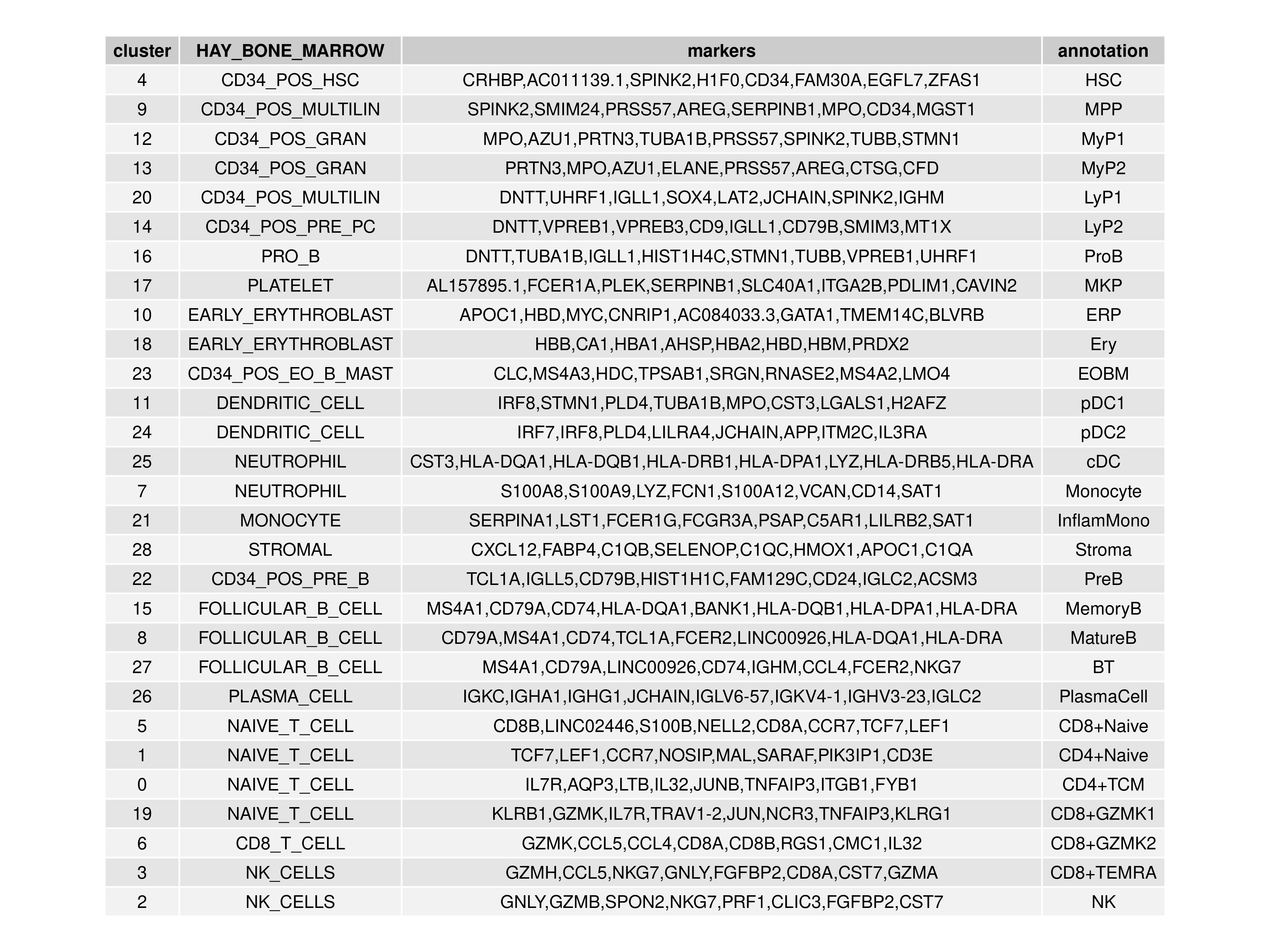


Figure 3.15: Annotation of the 29 clusters in the bone marrow dataset via marker genes and HAY\_BONE\_MARROW gene signature analysis.

**3.4.3 Visualising annotated clusters**

The final step is to relabel the cell clusters into the annotation labels. The resulting annotated cell types can then be visualised in t-SNE and UMAP projections (Figure @ref{clust-annotTsum}).

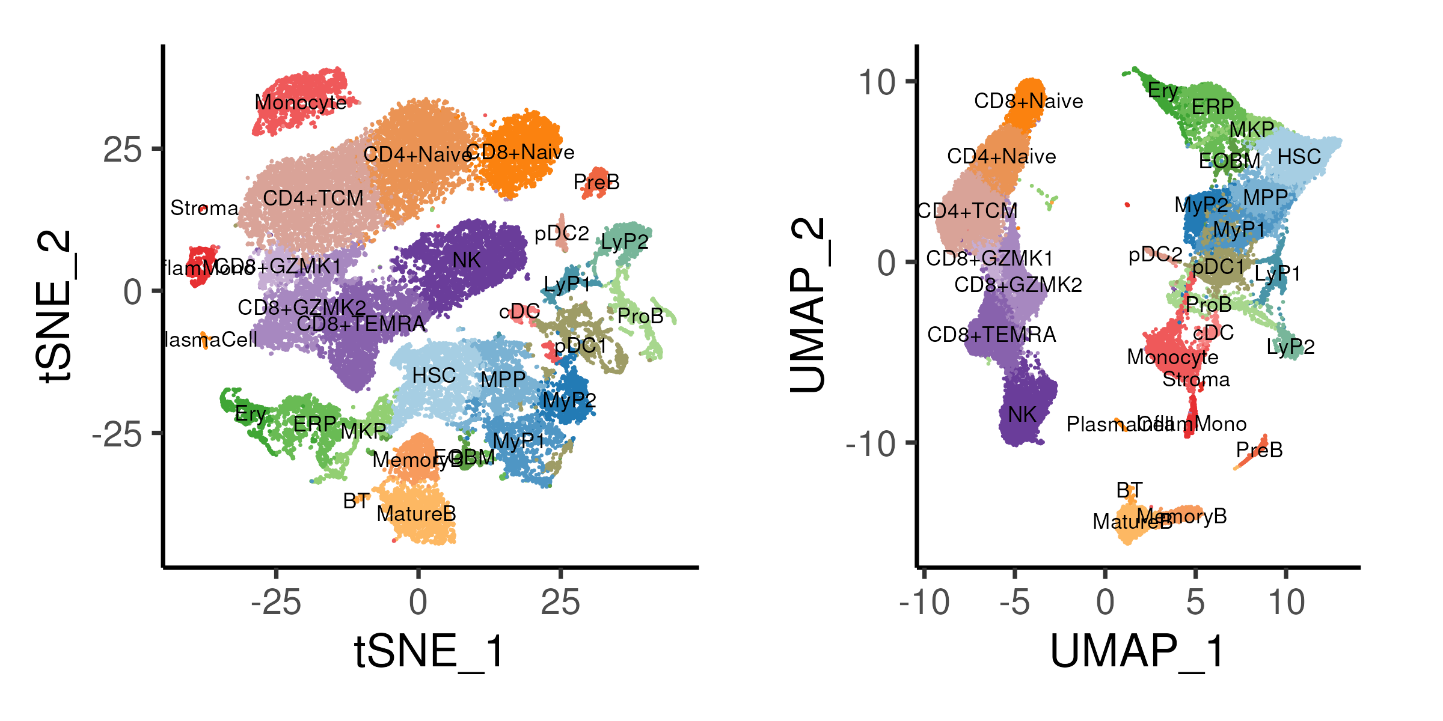


Figure 3.16: t-SNE and UMAP projections coloure by annotated cell types.

**3.4.4 Code**

Here, we provide code that extracts the most significant HAY\_BONE\_MARROW signature and top 8 marker genes by log-fold-change for each cluster to generate a table for the purposes of annotating the clusters. All 29 clusters are then annotated and a new metadata called celltype is being created in the Seurat object by mapping the cluster labels. The cell type label is then visualized on both t-SNE and UMAP projections.

*### D. Annotating clusters*

oupAnnot = data.table(cluster = reorderCluster)

*# Add in top 5 marker genes*

ggData <- oupMarker[, head(.SD, 8), by = "cluster"]

ggData <- ggData[, paste0(gene, collapse = ","), by = "cluster"]

colnames(ggData)[2] <- "markers"

oupAnnot <- ggData[oupAnnot, on = "cluster"]

*# Add in BONE\_MARROW enrichment*

ggData <- oupMarkerFunc[grep("BONE\_MARROW", ID)][, head(.SD, 1), by = "cluster"]

ggData <- ggData[, c("cluster", "ID")]

ggData$ID <- gsub("HAY\_BONE\_MARROW\_", "", ggData$ID)

colnames(ggData)[2] <- "HAY\_BONE\_MARROW"

oupAnnot <- ggData[oupAnnot, on = "cluster"]

*# Add in annotation*

oupAnnot$annotation <- c("HSC","MPP","MyP1","MyP2","LyP1","LyP2","ProB",

"ERP","Ery","MKP","EOBM",

"pDC1","pDC2","cDC","Monocyte","InflamMono","Stroma",

"PreB","MemoryB","MatureB","BT","PlasmaCell",

"CD8+Naive","CD4+Naive","CD4+TCM",

"CD8+GZMK1","CD8+GZMK2","CD8+TEMRA","NK")

*# Output table*

png("images/clustReAnnotTable.png",

width = 12, height = 9, units = "in", res = 300)

p1 <- tableGrob(oupAnnot, rows = NULL)

grid.arrange(p1)

dev.off()

*# Add new annotation into seurat and replot*

tmpMap <- oupAnnot$annotation

names(tmpMap) <- oupAnnot$cluster

seu$celltype <- tmpMap[as.character(seu$cluster)]

seu$celltype <- factor(seu$celltype, levels = tmpMap)

Idents(seu) <- seu$celltype *# Set seurat to use celltype*

*# Plot ideal resolution on tSNE and UMAP*

p1 <- DimPlot(seu, reduction = "tsne", pt.size = 0.1, label = TRUE,

label.size = 3, cols = colCls) + plotTheme + coord\_fixed()

p2 <- DimPlot(seu, reduction = "umap", pt.size = 0.1, label = TRUE,

label.size = 3, cols = colCls) + plotTheme + coord\_fixed()

ggsave(p1 + p2 & theme(legend.position = "none"),

width = 8, height = 4, filename = "images/clustReAnnotTsUm.png")

*# Save Seurat Object at end of each section*

saveRDS(seu, file = "bmSeu.rds")

**References**

Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders, and Ahmed Mahfouz. 2019. “A Comparison of Automatic Cell Identification Methods for Single-Cell RNA Sequencing Data.” *Genome Biology* 20: 194. <https://doi.org/10.1186/s13059-019-1795-z>.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.

Hay, Stuart B., Kyle Ferchen, Kashish Chetal, H. Leighton Grimes, and Nathan Salomonis. 2018. “The Human Cell Atlas Bone Marrow Single-Cell Interactive Web Portal.” *Experimental Hematology* 68 (December): 51–61. <https://doi.org/10.1016/j.exphem.2018.09.004>.

Kiselev, Vladimir Yu, Tallulah S. Andrews, and Martin Hemberg. 2019. “Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data.” *Nature Reviews Genetics* 20: 273–82. <https://doi.org/10.1038/s41576-018-0088-9>.

Kiselev, Vladimir Yu, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, et al. 2017. “SC3: Consensus Clustering of Single-Cell RNA-Seq Data.” *Nature Methods* 14: 483–86. <https://doi.org/10.1038/nmeth.4236>.

Nakamura-Ishizu, Ayako, Hitoshi Takizawa, and Toshio Suda. 2014. “The Analysis, Roles and Regulation of Quiescence in Hematopoietic Stem Cells.” *Development* 141 (24): 4656–66. <https://doi.org/10.1242/dev.106575>.

Pasquini, Giovanni, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. 2021. “Automated Methods for Cell Type Annotation on scRNA-Seq Data.” *Computational and Structural Biotechnology Journal* 19: 961–69. <https://doi.org/10.1016/j.csbj.2021.01.015>.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26: 139–40. <https://doi.org/10.1093/bioinformatics/btp616>.

Rozenblatt-Rosen, Orit, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. “The Human Cell Atlas: From Vision to Reality.” *Nature* 550 (7677): 451–53. <https://doi.org/10.1038/550451a>.

Schwartz, Gregory W., Yeqiao Zhou, Jelena Petrovic, Maria Fasolino, Lanwei Xu, Sydney M. Shaffer, Warren S. Pear, Golnaz Vahedi, and Robert B. Faryabi. 2020. “TooManyCells Identifies and Visualizes Relationships of Single-Cell Clades.” *Nature Methods* 17: 405–13. <https://doi.org/10.1038/s41592-020-0748-5>.

Soneson, Charlotte, and Mark D Robinson. 2018. “Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis.” *Nature Methods* 15: 255–61. <https://doi.org/10.1038/nmeth.4612>.

Squair, Jordan W., Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, et al. 2021. “Confronting False Discoveries in Single-Cell Differential Expression.” *Nature Communications* 12 (1). <https://doi.org/10.1038/s41467-021-25960-2>.

Svensson, Valentine. 2020. “Droplet scRNA-Seq Is Not Zero-Inflated.” *Nature Biotechnology* 38: 147–50. <https://doi.org/10.1038/s41587-019-0379-5>.

The Tabula Muris Consortium. 2018. “Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris.” *Nature* 562: 367–72. <https://doi.org/10.1038/s41586-018-0590-4>.

Tirosh, I., B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, et al. 2016. “Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq.” *Science* 352: 189–96. <https://doi.org/10.1126/science.aad0501>.

Zappia, Luke, and Alicia Oshlack. 2018. “Clustering Trees: A Visualization for Evaluating Clusterings at Multiple Resolutions.” *GigaScience* 7: giy083. <https://doi.org/10.1093/gigascience/giy083>.