Supplementary Document: validating BESCA single-cell signatures

Jitao David Zhang on behalf of the team

22 July, 2021

Contents

1 Motivation

The document describes the quality assessment of lists of genes preferentially expressed in individual cell types, referred to as *signatures* hereafter, offered by the BESCA software package.

2 Executive summary

We took two approaches. First, we compared BESCA signatures with tissue and cell-type signatures provided by BioQC, which are derived from bulk expression profiles. Second, we examined expression of BESCA signatures in the data collection of human protein atlas (HPA), which include other expression compendium including GTEx and FNATOM5 as well as newly generated data by HPA.

Both approaches revealed consistency between bulk and single-cell signatures. We also observed intriguing links between tissues and cell types that may lead to new biological insights.

3 Analysis

3.1 Comparing BESCA and BioQC signatures

We first download the BESCA signatures from the project's repository and parse its content.

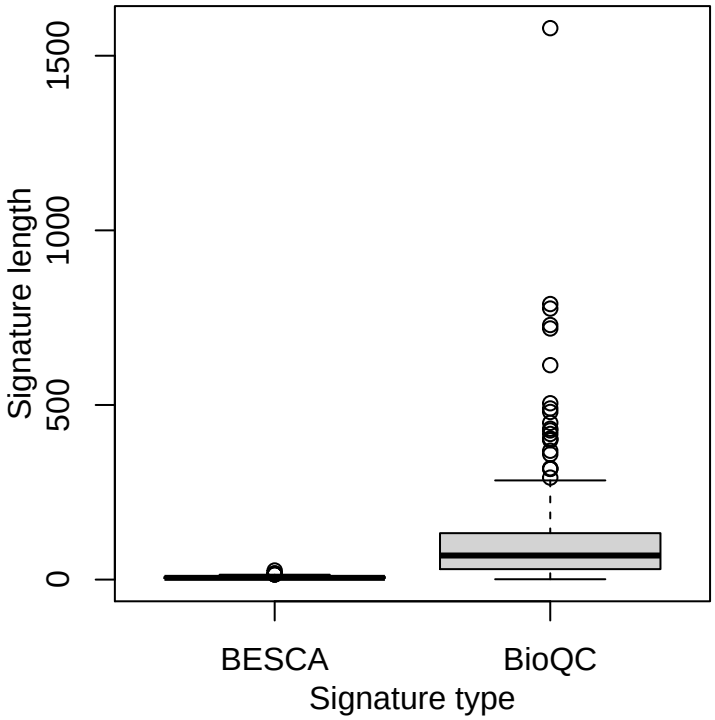
```
bescaGmt <- "data/20210719-BESCA-CellNames_scseqCMs6_sigs.gmt"
if(!file.exists(bescaGmt)) {
  url <- "https://raw.githubusercontent.com/bedapub/besca/master/besca/datasets/genesets/CellNames_scseq
  download.file(url, destfile=bescaGmt)
}
bescaSignature <- BioQC::readGmt(bescaGmt)
```

```
## Warning in readLines(filename): incomplete final line found on 'data/20210719-
## BESCA-CellNames_scseqCMs6_sigs.gmt'
```

Next we parse the BioQC signatures.

```
bioqcSig <- BioQC::readCurrentSignatures()
```

```
bescaSigLen <- gsGeneCount(bescaSignature)
bioqcSigLen <- gsGeneCount(bioqcSig)
{
  compactPar()
  boxplot(list(BESCA=bescaSigLen, BioQC=bioqcSigLen),
          xlab="Signature type", ylab="Signature length")
}
```



Notice that BESCA signatures derived from single-cell studies are much shorter than BioQC signatures derived from bulk gene expression profiles. To assess the consistency between the signatures, we calculate pairwise overlap coefficients between two types of signatures. The overlap coefficient of two sets A and B is defined as

$$\text{overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

If set A is a subset of B (or *vice versa*), the overlap coefficient is equal to 1.

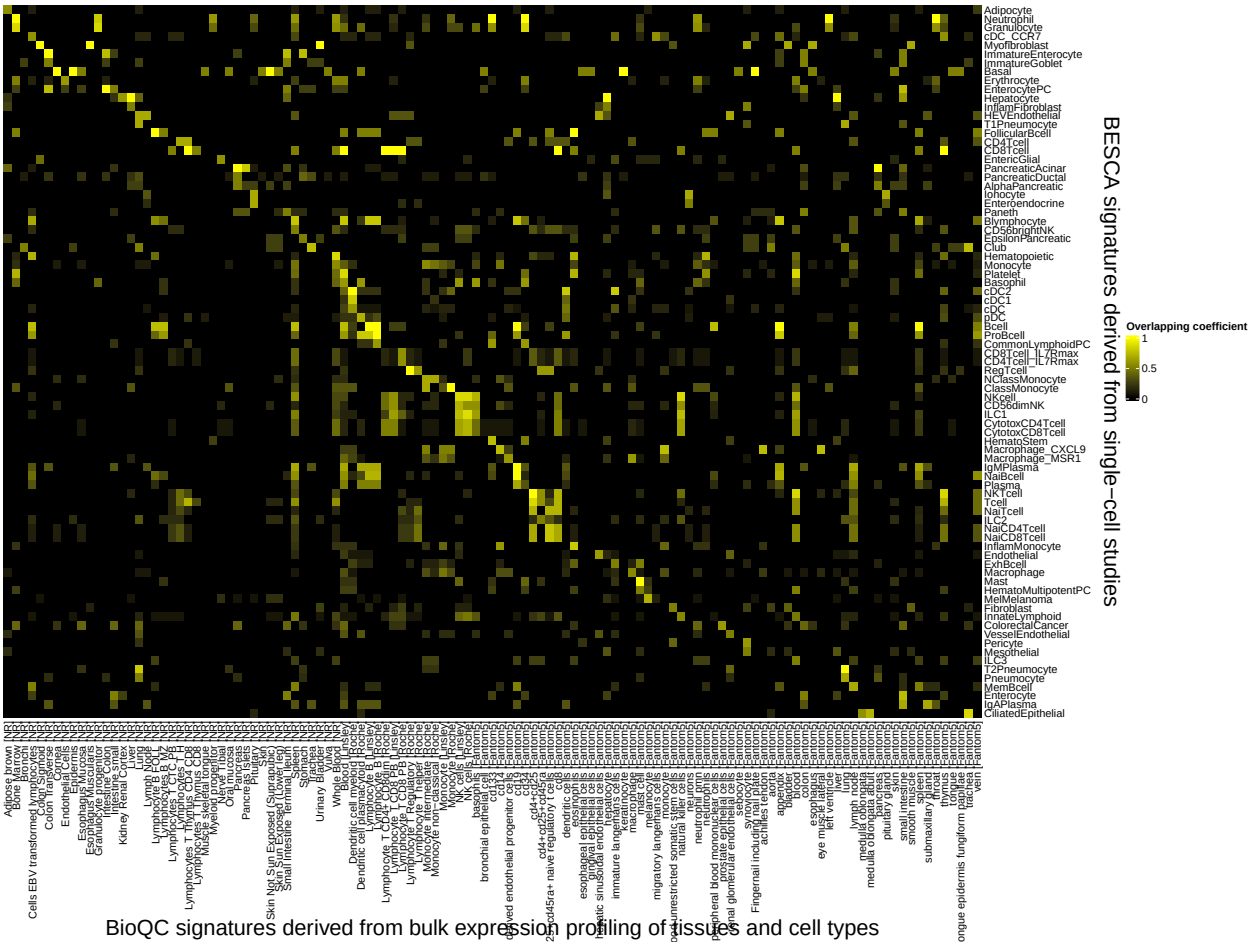
```
bescaBioQCOverlap <- sapply(bioqcSig, function(bq) {
  sapply(bescaSignature, function(be) {
    overlapCoefficient(be$genes, bq$genes)
  })
})
```

We consider either single-cell or bulk signatures that have at least one signature of the other type with the overlap coefficient equal to or larger than 0.5.

```
isBescaStrOverlap <- apply(bescaBioQCOverlap, 1, function(x) any(x>=0.5))
isBioqcStrOverlap <- apply(bescaBioQCOverlap, 2, function(x) any(x>=0.5))
bescaBioQCstrOverlap <- bescaBioQCOverlap[isBescaStrOverlap, isBioqcStrOverlap]
bbCasord <- cascadeOrder(bescaBioQCstrOverlap)
bbCasOverlap <- bescaBioQCstrOverlap[bbCasord,]
colnames(bbCasOverlap) <- prettySigNames(colnames(bbCasOverlap))
```

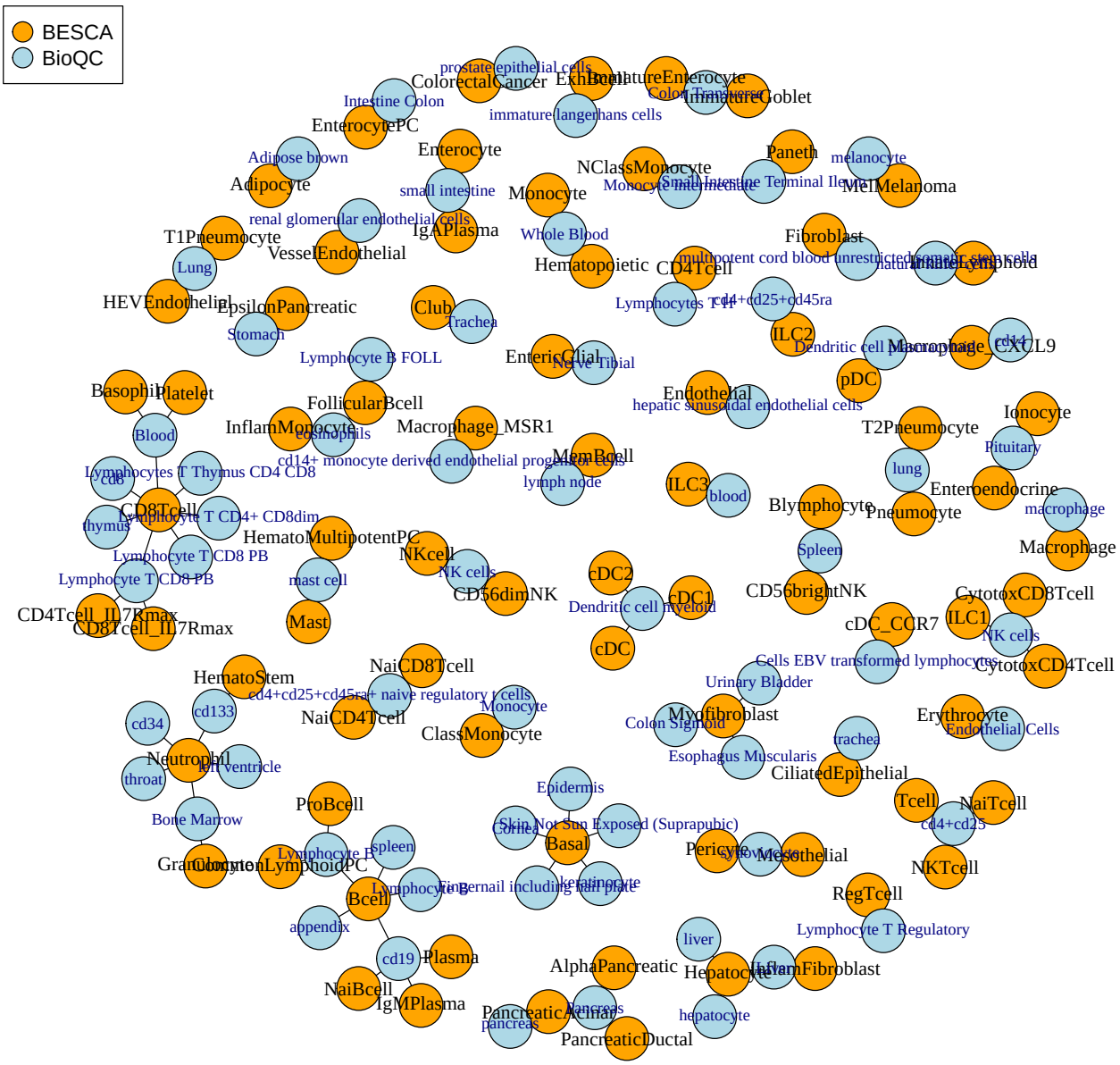
The haetmap below visualizes all pairs of signatures.

```
ComplexHeatmap::Heatmap(bbCasOverlap,
                        name="Overlapping coefficient",
                        cluster_columns=FALSE, cluster_rows = FALSE,
                        col=circlize::colorRamp2(breaks=c(0, 0.5, 1),
                                 colors=blackyellow(3)),
                        column_title = paste("BioQC signatures derived from",
                        "bulk expression profiling of tissues and cell types"),
                        row_title = "BESCA signatures derived from single-cell studies",
                        column_title_side = "bottom", column_title_gp = gpar(fontsize=20),
                        row_title_side = "right", row_title_gp = gpar(fontsize=20),
                        row_names_gp = gpar(fontsize=10),
                        column_names_gp = gpar(fontsize=10))
```



To assist visual inspection, we convert the matrix into a bipartite graph. For each single-cell signature in BESCA, we report all bulk signatures in BioQC that have either the highest, non-zero overlap coefficient among all signatures, or have the overlap coefficient equal to 1. The bipartite graph is visualized below.

```
source("2021-07-toribios.R")
bbCasTops <- applyTopOrIncAndNotExclFilter(bbCasOverlap, MARGIN=1, top=1,
                                           falseValue=0,
                                           incFunc=function(x) x==1,
                                           excFunc=function(x) x==0)
bbCasUsedTops <- removeColumns(bbCasTops, function(x) !any(x>0))
bbGraph <- buildBescaIncidenceGraph(bbCasUsedTops)
bbLayout <- layout_with_fr(bbGraph, niter=2000)
{
  plot(bbGraph, layout=bbLayout)
  legend('topleft',legend=c("BESCA", "BioQC"), pch=21, pt.bg=c("orange", "lightblue"),
         cex=1.2, pt.cex=2.5)
}
```



The graph represents (1) gene expression signatures derived from single-cell sequencing studies extracted and offered by BESCA (orange nodes), (2) gene expression signatures derived from bulk gene-expression profiling studies performed with microarray and next-generation sequencing extracted and offered by BioQC (blue nodes), and (3) their similarities encoded in edges.

We manually inspected the signatures. Despite different vocabularies were used in bulk and single-cell studies, most cell/tissue types match by corresponding biological entities. Examples include dendritic cell myeloid (bulk) and cDC/cDC1/cDC2 (single-cell), trachea (bulk) and cilated epithelial (single-cell), and liver and isolated hepatocytes (bulk) and hepatocytes (single-cell). They suggest that the signatures provided by BESCA and BioQC are well consistent with each other.

At the same time, there are a few intriguing links, for instance multipotent cord blood unrestricted somatic stem cells (bulk) and fibroblasts (single-cell), and pituitary (bulk) and enteroendocrine and ionocyte (single-cell). Several explanations can be biologically plausible. For example, fibroblasts show heterogeneous, context-dependent expression patterns (Buechler, *et al.* "Cross-Tissue Organization of the Fibroblast Lineage." Nature (2021)). Enteroendocrine cells are sensory cells of the gut that communicate by releasing hormones locally, which fulfills a similar endocrine function of the pituitary gland. Two genes, `CHGA` and `ASCL1`, are

shared between bulk pituitary signature and single-cell ionocyte signatures and both are preferentially expressed by neuroendocrine cells. We failed to identify studies reporting whether they have identical or distinct biological functions in these contexts. It suggests that integrative analysis of single-cell and bulk gene expression signatures using *BESCA* and *BioQC* reveals hidden patterns that warrant further research.

The code below shows intersection between single-cell signatures of ionocytes and bulk signatures of pituitary gland.

```
intersect(bescaSignature[["Ionocyte"]]$genes,
          bioqcSig[["pituitary_gland_Fantom5_Tissue_0.7_3"]]$genes)
```

```
## [1] "CHGA"  "ASCL1"
```

3.2 Expression of single-cell signatures in Human Protein Atlas.

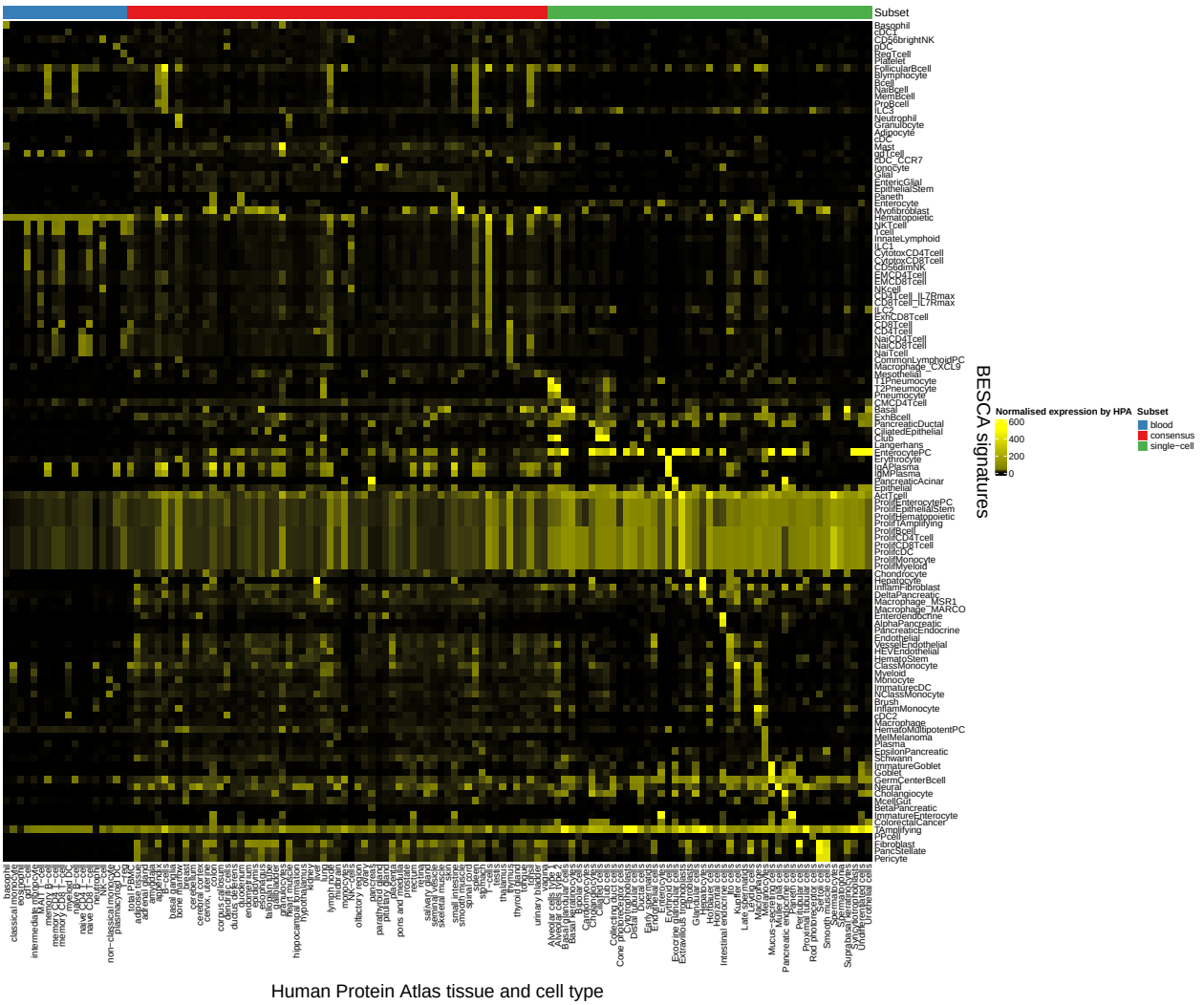
Besides checking consistency of single-cell and bulk signatures, we also directly examined expression of single-cell signatures in bulk and single-cell studies deposited in publicly available gene expression compendia. For this purpose, we used consensus expression data, the Blood Atlas, and the single-cell expression data collected by the Human Protein Atlas (HPA) project.

```
bescaSigHPAfile <- "data/bescaSigHPA.tsv.gz"
if(!file.exists(bescaSigHPAfile)) {
  consensus <- read_tsv("data/rna_consensus.tsv", col_types = "cccn") %>%
    mutate(Subset="consensus", GeneSymbol=`Gene name`, TC=Tissue)
  blood <- read_tsv("data/rna_blood_cell.tsv", col_types="cccnnn") %>%
    mutate(Subset="blood", GeneSymbol=`Gene name`, TC=`Blood cell`)
  hpasc <- read_tsv("data/rna_single_cell_type.tsv", col_types="cccn") %>%
    mutate(Subset="single-cell", GeneSymbol=`Gene name`, TC=`Cell type`)

  bescaSigGenes <- munion(gsGenes(bescaSignature))
  selTbl <- function(x) x %>% filter(GeneSymbol %in% bescaSigGenes) %>%
    dplyr::select(Subset, GeneSymbol, TC, NX)
  bescaSigHPA <- rbind(consensus %>% selTbl,
                       blood %>% selTbl,
                       hpasc %>% selTbl)
  write_tsv(bescaSigHPA, bescaSigHPAfile)
} else {
  bescaSigHPA <- read_tsv(bescaSigHPAfile, col_type="cccn")
}
bescaDf <- list2df(gsGenes(bescaSignature), col.names = c("GeneSet", "GeneSymbol"))
bescaSigMat <- inner_join(bescaSigHPA, bescaDf, by="GeneSymbol") %>%
  group_by(Subset, GeneSet, TC) %>%
  summarise(MedianNX=median(NX), .groups="drop") %>%
  longdf2matrix(., row.col="GeneSet", column.col="TC", value.col="MedianNX")
bescaSigMatOrd <- cascadeOrder(bescaSigMat)
bescaSigMatOrdered <- bescaSigMat[bescaSigMatOrd, ]
hpaSubset <- matchColumn(colnames(bescaSigMat), bescaSigHPA, "TC")$Subset
hpaSubsetCol <- RColorBrewer::brewer.pal(3, "Set1")
bescaSigMatAnno <- HeatmapAnnotation(Subset=factor(hpaSubset),
                                     col = list(Subset=c("consensus"=hpaSubsetCol[1],
                                                         "blood"=hpaSubsetCol[2],
                                                         "single-cell"=hpaSubsetCol[3])),
                                     which = "column",
                                     annotation_name_side = "right")
```

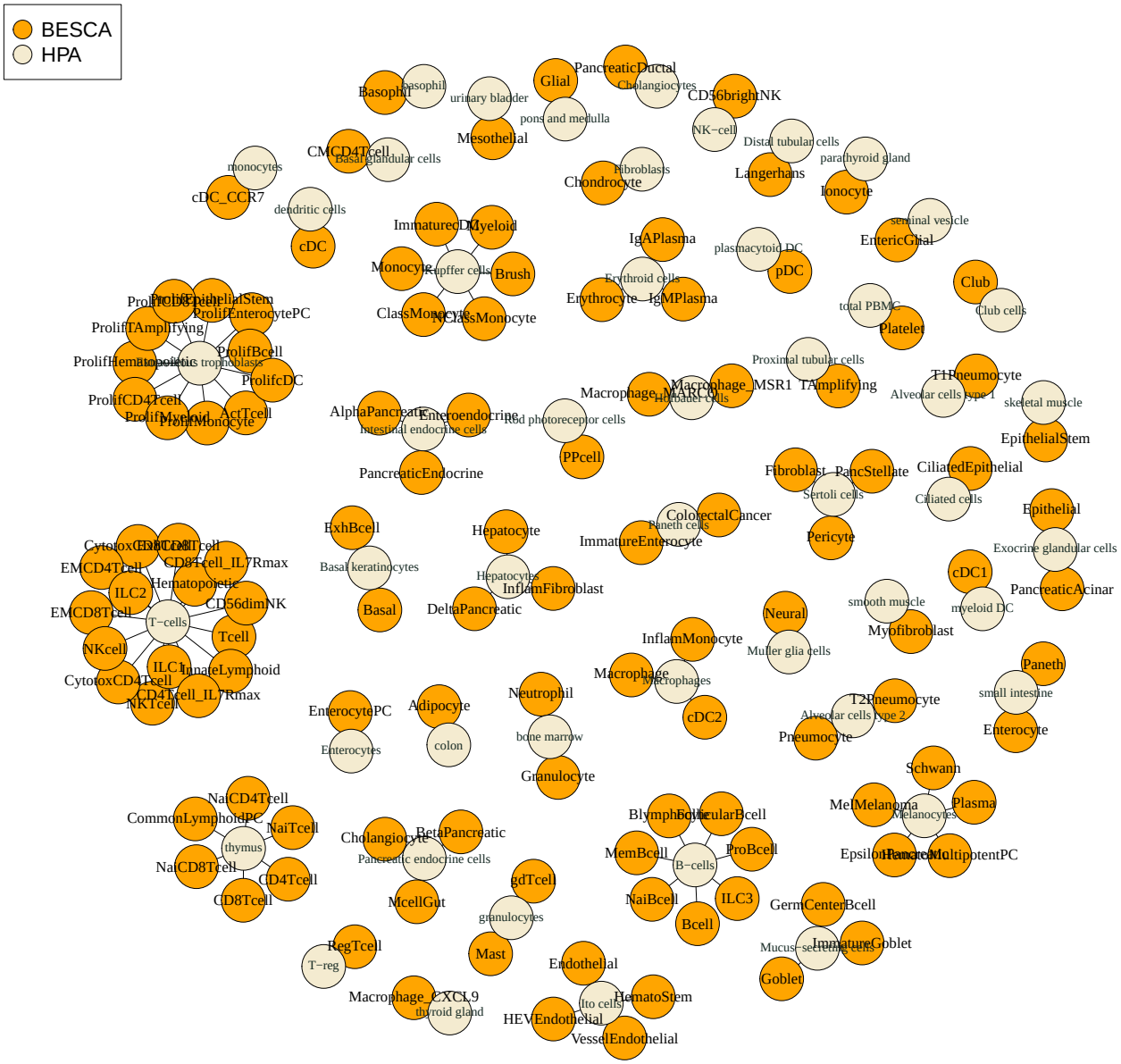
The heatmap belwo shows the expression of BESCA signatures in HPA. Each signature is represented by the median *consensus, normalized expression* (NX) of its genes. HPA datasets are colored by the subset.

```
ComplexHeatmap::Heatmap(bescaSigMatOrdered,
                        top_annotation = bescaSigMatAnno,
                        name="Normalised expression by HPA",
                        cluster_columns=FALSE, cluster_rows = FALSE,
                        col=circlize::colorRamp2(breaks=c(0,
                                                          quantile(bescaSigHPA$NX, 0.9),
                                                          quantile(bescaSigHPA$NX, 0.99)),
                                colors=blackyellow(3)),
                        column_title = "Human Protein Atlas tissue and cell type",
                        row_title = "BESCA signatures",
                        column_title_side = "bottom", column_title_gp = gpar(fontsize=20),
                        row_title_side = "right", row_title_gp = gpar(fontsize=20),
                        row_names_gp = gpar(fontsize=10),
                        column_names_gp = gpar(fontsize=10))
```



Similar to what we did for the comparison between BESCA and BioQC signatures, we build a bipartite graph of BESCA signatures and HPA tissue and cell types. For each single-cell signature in BESCA, we report all HPA tissue and cell types that have the highest, non-zero NX.

```
hpaCol <- "#f4ebd0" ## cream
bhTops <- applyTopOrIncAndNotExclFilter(bescaSigMatOrdered, MARGIN=1, top=1,
                                        falseValue=0,
                                        excFunc=function(x) x==0)
bhUsedTops <- removeColumns(bhTops, function(x) !any(x>0))
bhGraph <- buildBescaIncidenceGraph(bhUsedTops,
                                    column_node_color = hpaCol,
                                    column_label_color = "#122620") ## charcoal
bhLayout <- layout_with_fr(bhGraph, niter=2000)
{
  plot(bhGraph, layout=bhLayout)
  legend('topleft',legend=c("BESCA", "HPA"), pch=21, pt.bg=c("orange", hpaCol),
         cex=1.5, pt.cex=3)
}
```



We inspected the graph manually and found good consistency between BESCA signatures and HPA tissue

and cell types. Similar to the comparison of BESCA and BioQC signatures, we found some intriguing links that deserve further interpretation and research.

3.3 Exporting results

We export the matrices and graphs so that other visualization and analysis methods can be applied to them.

```
write_gct(bbCasOverlap, "data/Besca-BioQC-signature-overlapping-coefficients.gct")
write_gct(bbCasOverlap, "data/Besca-signature-median-NX-in-HPA.gct")
igraph::write_graph(bbGraph, "data/Besca-BioQC-bbGraph.graphml", format="graphml")
igraph::write_graph(bhGraph, "data/Besca-HPA-bhGraph.graphml", format="graphml")
```

4 Conclusions

We observe considerable consistency between single-cell signatures provided by BESCA and existing data and knowledge of tissue- and cell-type-specific gene expression.

5 Acknowledgment

I think colleagues of the BEDA team for continuous support and dicsussions.

6 Session information

```
bedaInfo()
```

```
## A Pharmaceutical Sciences (PS) Bioinformatics and Exploratory Data Analysis (BEDA) project
##
## [pstore path]
##    /mnt/projects/2021-07-BESCArevision
## [URL]
##    /mnt/projects/2021-07-BESCArevision
## [git]
##    git@github.roche.com:BEDA-recipes/2021-07-BESCArevision-BioQC.git
## [User]R
##       R
##    david
```

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Linux Mint 19.3
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.2.20.so
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
```

```
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] igraph_1.2.6         ComplexHeatmap_2.6.2 BioQC_1.21.2
##  [4] Biobase_2.50.0       BiocGenerics_0.36.0  httr_1.4.2
##  [7] forcats_0.5.1        stringr_1.4.0        dplyr_1.0.6
## [10] purrr_0.3.4          readr_1.4.0          tidyr_1.1.3
## [13] tibble_3.1.2         ggplot2_3.3.4        tidyverse_1.3.0
## [16] openxlsx_4.2.3       ribiosGSEA_1.5-3     ribiosPlot_1.2-9
## [19] ribiosIO_1.0-62      ribiosUtils_1.5-14
##
## loaded via a namespace (and not attached):
##   [1] readxl_1.3.1                backports_1.2.1
##   [3] circlize_0.4.12             ribiosExpression_1.1-1
##   [5] splines_4.0.5               BiocParallel_1.24.1
##   [7] GenomeInfoDb_1.26.5         made4_1.64.0
##   [9] sva_3.38.0                  digest_0.6.27
##  [11] htmltools_0.5.1.1           GO.db_3.12.1
##  [13] fansi_0.5.0                 magrittr_2.0.1
##  [15] memoise_2.0.0               cluster_2.1.1
##  [17] limma_3.46.0                Biostrings_2.58.0
##  [19] annotate_1.68.0             modelr_0.1.8
##  [21] matrixStats_0.58.0          prettyunits_1.1.1
##  [23] colorspace_2.0-1            blob_1.2.1
##  [25] rvest_1.0.0                 haven_2.3.1
##  [27] xfun_0.23                   crayon_1.4.1
##  [29] RCurl_1.98-1.3              jsonlite_1.7.2
##  [31] graph_1.68.0                genefilter_1.72.1
##  [33] survival_3.2-10             glue_1.4.2
##  [35] gtable_0.3.0                zlibbioc_1.36.0
##  [37] XVector_0.30.0              GetoptLong_1.0.5
##  [39] DelayedArray_0.16.3         ribiosArg_1.4-0
##  [41] shape_1.4.5                 scales_1.1.1
##  [43] vsn_3.58.0                  DBI_1.1.1
##  [45] edgeR_3.32.1                Rcpp_1.0.6
##  [47] xtable_1.8-4                progress_1.2.2
##  [49] gage_2.40.2                 clue_0.3-58
##  [51] bit_4.0.4                   preprocessCore_1.52.1
##  [53] stats4_4.0.5                gplots_3.1.1
##  [55] RColorBrewer_1.1-2          ellipsis_0.3.2
##  [57] pkgconfig_2.0.3             XML_3.99-0.6
##  [59] dbplyr_2.1.1                locfit_1.5-9.4
##  [61] utf8_1.2.1                  tidyselect_1.1.1
##  [63] rlang_0.4.11                AnnotationDbi_1.52.0
##  [65] munsell_0.5.0               cellranger_1.1.0
##  [67] tools_4.0.5                 cachem_1.0.4
```

```
##  [69] cli_2.5.0                  generics_0.1.0
##  [71] RSQLite_2.2.5              ade4_1.7-16
##  [73] broom_0.7.6                evaluate_0.14
##  [75] fastmap_1.1.0              yaml_2.2.1
##  [77] knitr_1.33                 bit64_4.0.5
##  [79] fs_1.5.0                   zip_2.1.1
##  [81] caTools_1.18.2             KEGGREST_1.30.1
##  [83] nlme_3.1-152               xml2_1.3.2
##  [85] ribiosAnnotation_3.4-3     compiler_4.0.5
##  [87] rstudioapi_0.13            png_0.1-7
##  [89] affyio_1.60.0              reprex_2.0.0
##  [91] ribiosNGS_1.1-23           stringi_1.6.2
##  [93] highr_0.9                  lattice_0.20-41
##  [95] Matrix_1.3-2               vctrs_0.3.8
##  [97] pillar_1.6.1               lifecycle_1.0.0
##  [99] BiocManager_1.30.12        GlobalOptions_0.1.2
## [101] bitops_1.0-6               GenomicRanges_1.42.0
## [103] R6_2.5.0                   affy_1.68.0
## [105] KernSmooth_2.23-18         IRanges_2.24.1
## [107] MASS_7.3-53.1              gtools_3.8.2
## [109] assertthat_0.2.1           SummarizedExperiment_1.20.0
## [111] rjson_0.2.20               withr_2.4.2
## [113] S4Vectors_0.28.1           GenomeInfoDbData_1.2.4
## [115] mgcv_1.8-34                hms_1.1.0
## [117] rmarkdown_2.8              MatrixGenerics_1.2.1
## [119] Cairo_1.5-12.2             scatterplot3d_0.3-41
## [121] lubridate_1.7.10
```