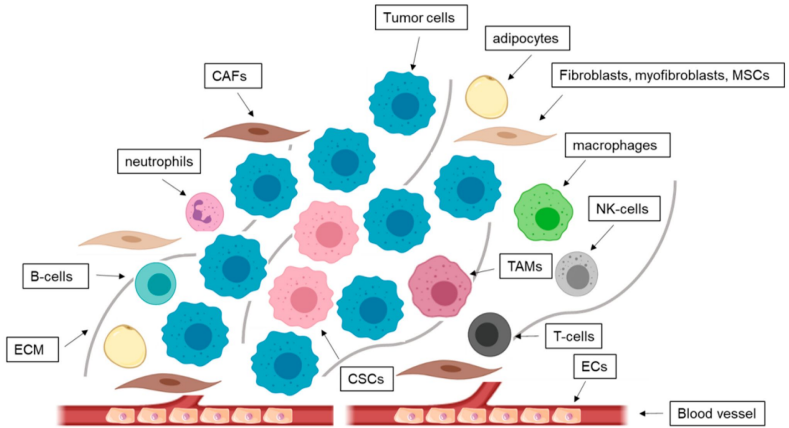
Exploring Human Breast Cancer Atlas scRNA-Seq Data

Team 4
Arthi Hariharan, Seowoo Kim, Taeim Kwon

Biological Concepts behind Questions of Interest

- Tumor microenvironment is heterogenous
  - Immune Cells - T cells, B cells
  - Mesenchymal Stem cells
  - FIbroblasts - Normal, Cancerous
- Breast Cancer Subtypes - HER2+, ER+, TNBC
  - Different Gene Signatures
  - Treatment resistance
  - Gene expression profile - Prognosis



Adapted from Heesch A et al., 2020

Biological Questions:

- Identify the differential expression of genes across breast cancer subtypes.
  - Focus on cancer hallmarks - Metastasis and Inflammation markers (CXCR family) and Angiogenesis markers (VEGF,PDGF family) .

- Show a gene co-expression network in the cells involved with tumor microenvironment, particularly for mesenchymal cells.

Introduction to Dataset

- GSE176078_Wu_etal_2021_BRCA_scRNASeq.tar.gz:
  - A single-cell and spatially resolved atlas of human breast cancers by Wu et al.
  - scRNA-Seq on 26 primary tumors (11 ER+, 5 HER2+ and 10 TNBC)
  - Files
    - Count_matrix_barcodes.tsv
    - Count_matrix_genes.tsv
    - Count_matrix_sparse.mtx
    - Metadata.csv
  - 29,733 features and 100,064 samples

Initial Processing

```
> project
An object of class Seurat
29733 features across 100064 samples within 1 assay
Active assay: RNA (29733 features, 0 variable features)
```

1,777 samples were filtered out

```
> project
An object of class Seurat
29733 features across 98287 samples within 1 assay
Active assay: RNA (29733 features, 2000 variable features)
 2 dimensional reductions calculated: pca, umap
```

Set up Seurat

1. Read in the files
2. Create the Seurat object

Pre-processing

1. Quality control
2. Normalize the data
3. Select the highly variable features
4. Scale the data

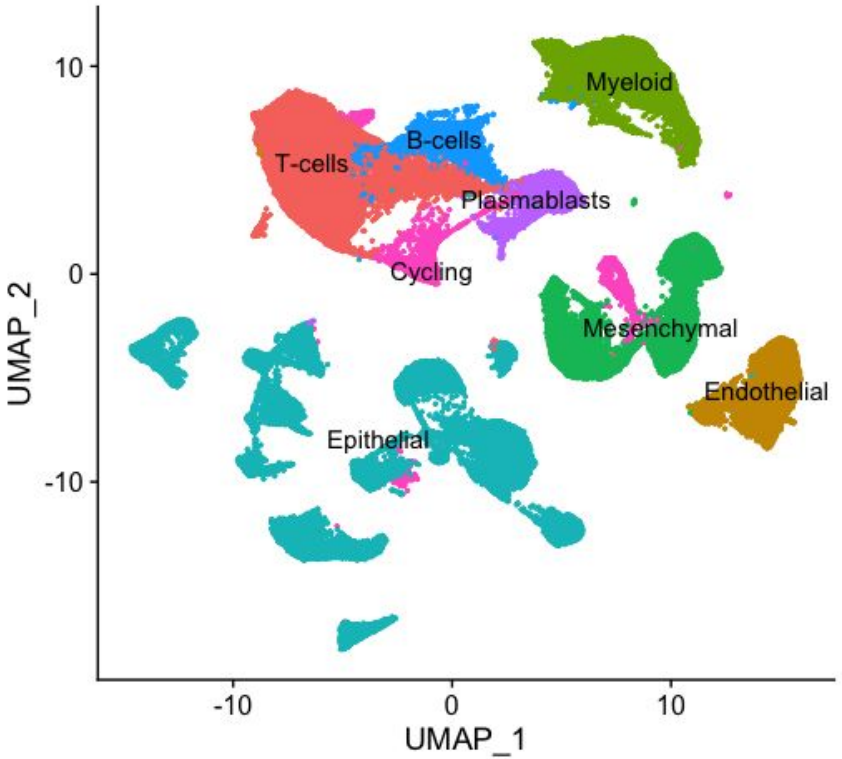
Dimensionality

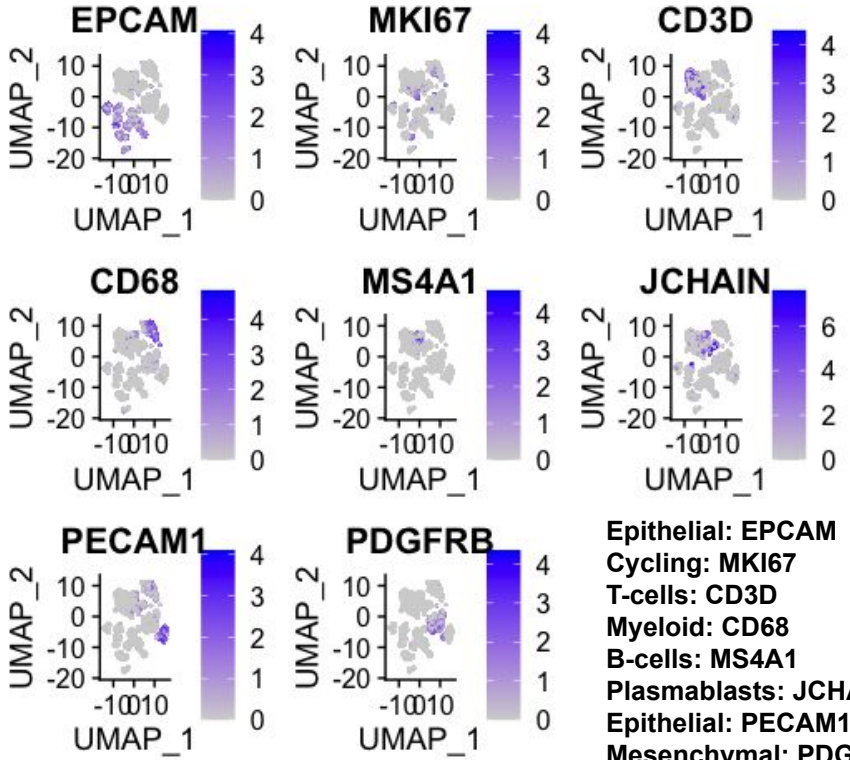
1. Run PCA
2. Check an ‘elbow plot’ to choose the dimension of the data
3. Run ‘UMAP’

Cell Type Annotation

1. Automatic annotation with the metadata
2. Manual annotation by finding marker genes from each cluster

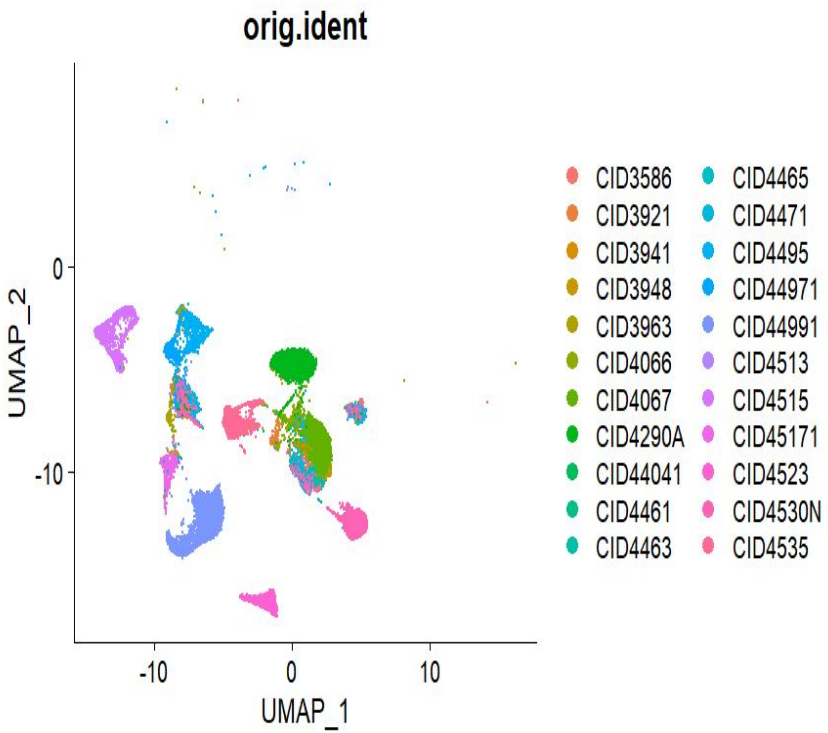
Reproduced outputs

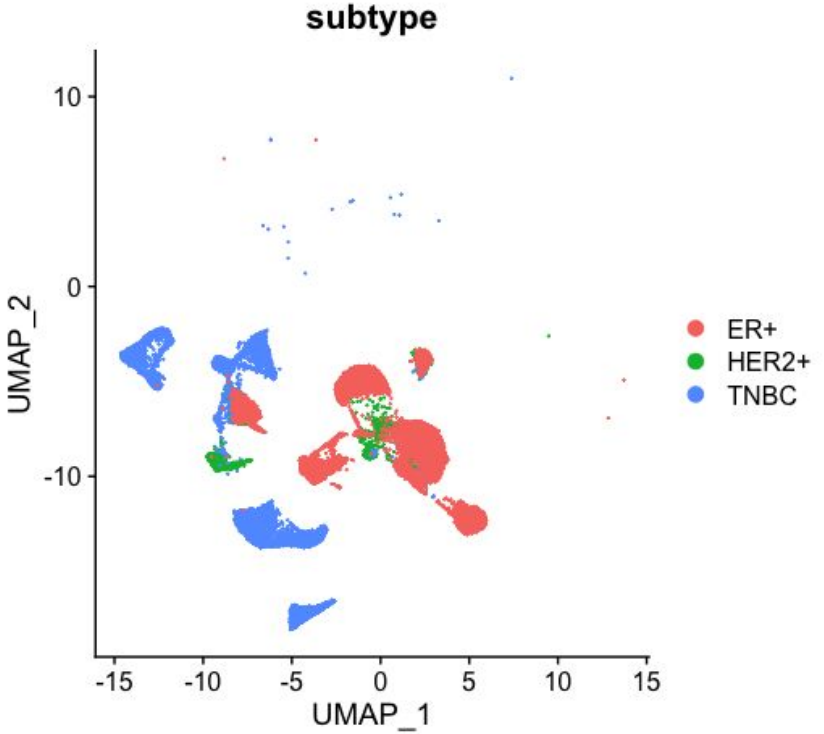




Epithelial: EPCAM
Cycling: MKI67
T-cells: CD3D
Myeloid: CD68
B-cells: MS4A1
Plasmablasts: JCHAIN
Epithelial: PECAM1
Mesenchymal: PDGFRB

Reproduced outputs

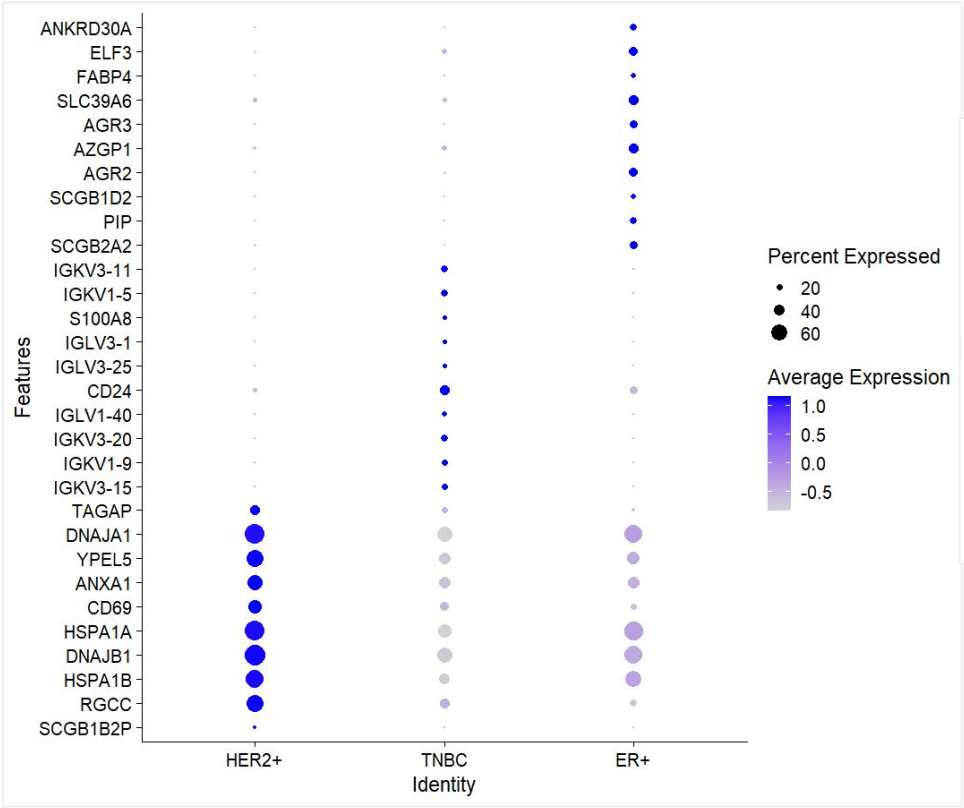


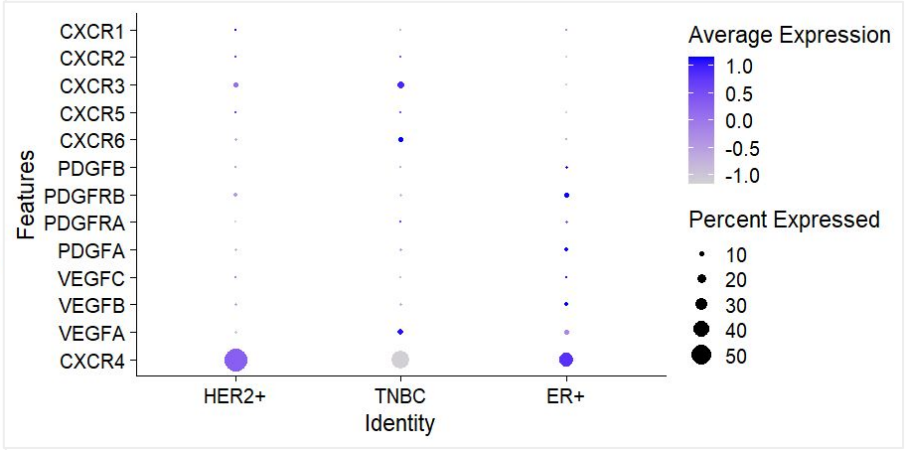


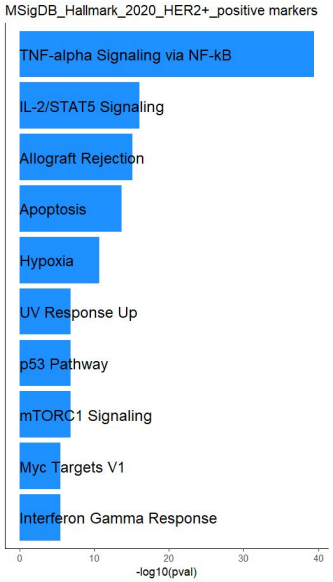
Differential Expression Analysis

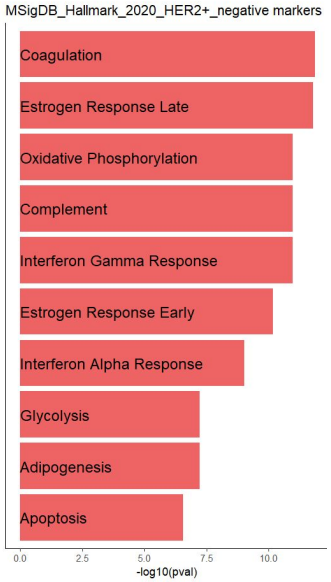
- Differential Expression Analysis
  - “FindMarkers” function used to find differentially expressed genes
  - Highly expressed markers in HER2+, ER+ and TNBC
  - CXCR family genes, VEGF, PDGF genes across subtypes
  - EnrichR using MSigDb_hallmarks to find upregulated or downregulated GO terms - “DEEnrichRPlot”

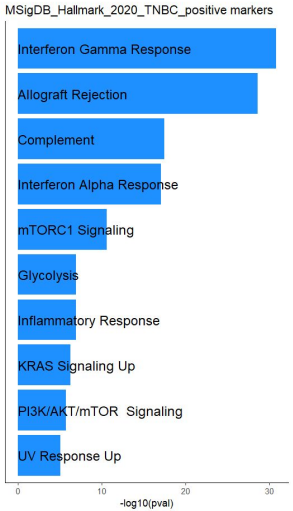
Differential Expression Analysis

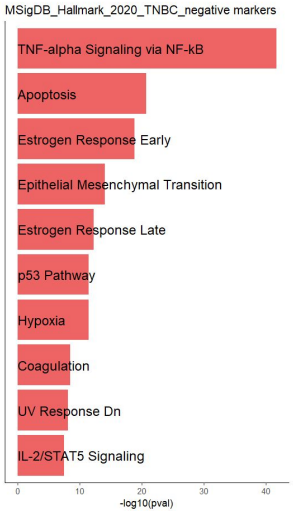


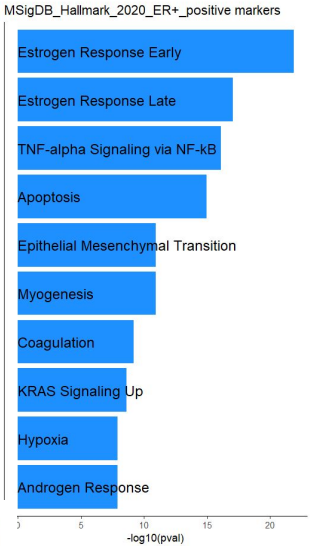


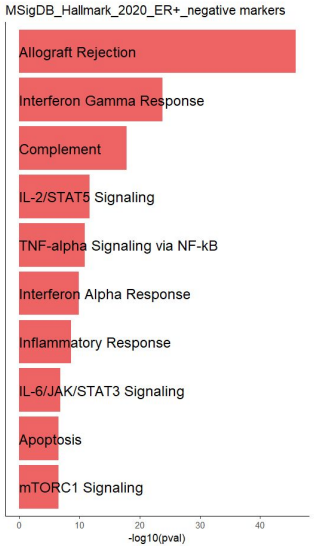

MSigDB_Hallmark_2020_HER2+_positive markers
TNF-alpha Signaling via NF-kB
IL-2/STAT5 Signaling
Allograft Rejection
Apoptosis
Hypoxia
UV Response Up
p53 Pathway
mTORC1 Signaling
Myc Targets V1
Interferon Gamma Response
0
10
20
30
40
-log10(pval)


MSigDB_Hallmark_2020_HER2+_negative markers
Coagulation
Estrogen Response Late
Oxidative Phosphorylation
Complement
Interferon Gamma Response
Estrogen Response Early
Interferon Alpha Response
Glycolysis
Adipogenesis
Apoptosis
0.0
2.5
5.0
7.5
10.0
-log10(pval)


MSigDB_Hallmark_2020_TNBC_positive markers
Interferon Gamma Response
Allograft Rejection
Complement
Interferon Alpha Response
mTORC1 Signaling
Glycolysis
Inflammatory Response
KRAS Signaling Up
PI3K/AKT/mTOR Signaling
UV Response Up
0
10
20
30
-log10(pval)


MSigDB_Hallmark_2020_TNBC_negative markers
TNF-alpha Signaling via NF-kB
Apoptosis
Estrogen Response Early
Epithelial Mesenchymal Transition
Estrogen Response Late
p53 Pathway
Hypoxia
Coagulation
UV Response Dn
IL-2/STAT5 Signaling
0
10
20
30
40
-log10(pval)

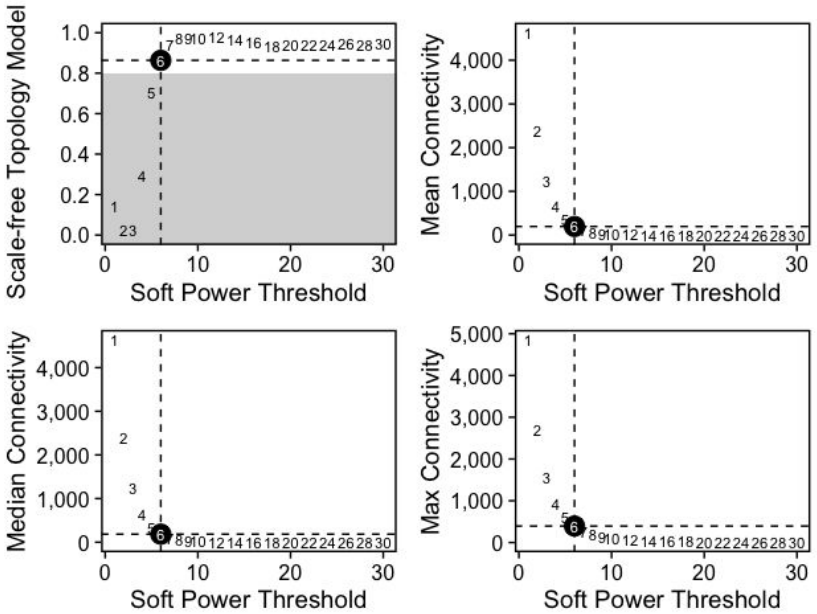

MSigDB_Hallmark_2020_ER+_positive markers
Estrogen Response Early
Estrogen Response Late
TNF-alpha Signaling via NF-kB
Apoptosis
Epithelial Mesenchymal Transition
Myogenesis
Coagulation
KRAS Signaling Up
Hypoxia
Androgen Response
0
5
10
15
20
-log10(pval)

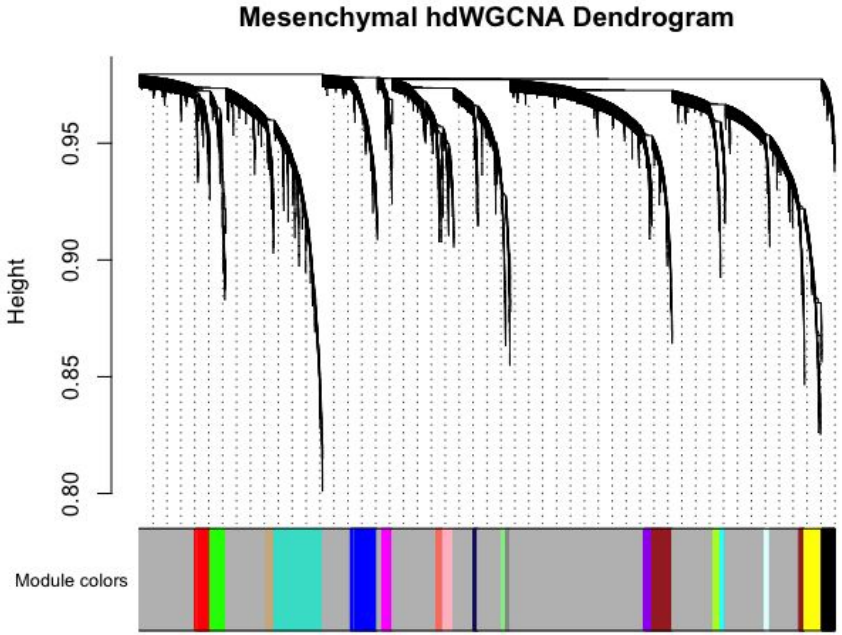

MSigDB_Hallmark_2020_ER+_negative markers
Allograft Rejection
Interferon Gamma Response
Complement
IL-2/STAT5 Signaling
TNF-alpha Signaling via NF-kB
Interferon Alpha Response
Inflammatory Response
IL-6/JAK/STAT3 Signaling
Apoptosis
mTORC1 Signaling
0
10
20
30
40
-log10(pval)

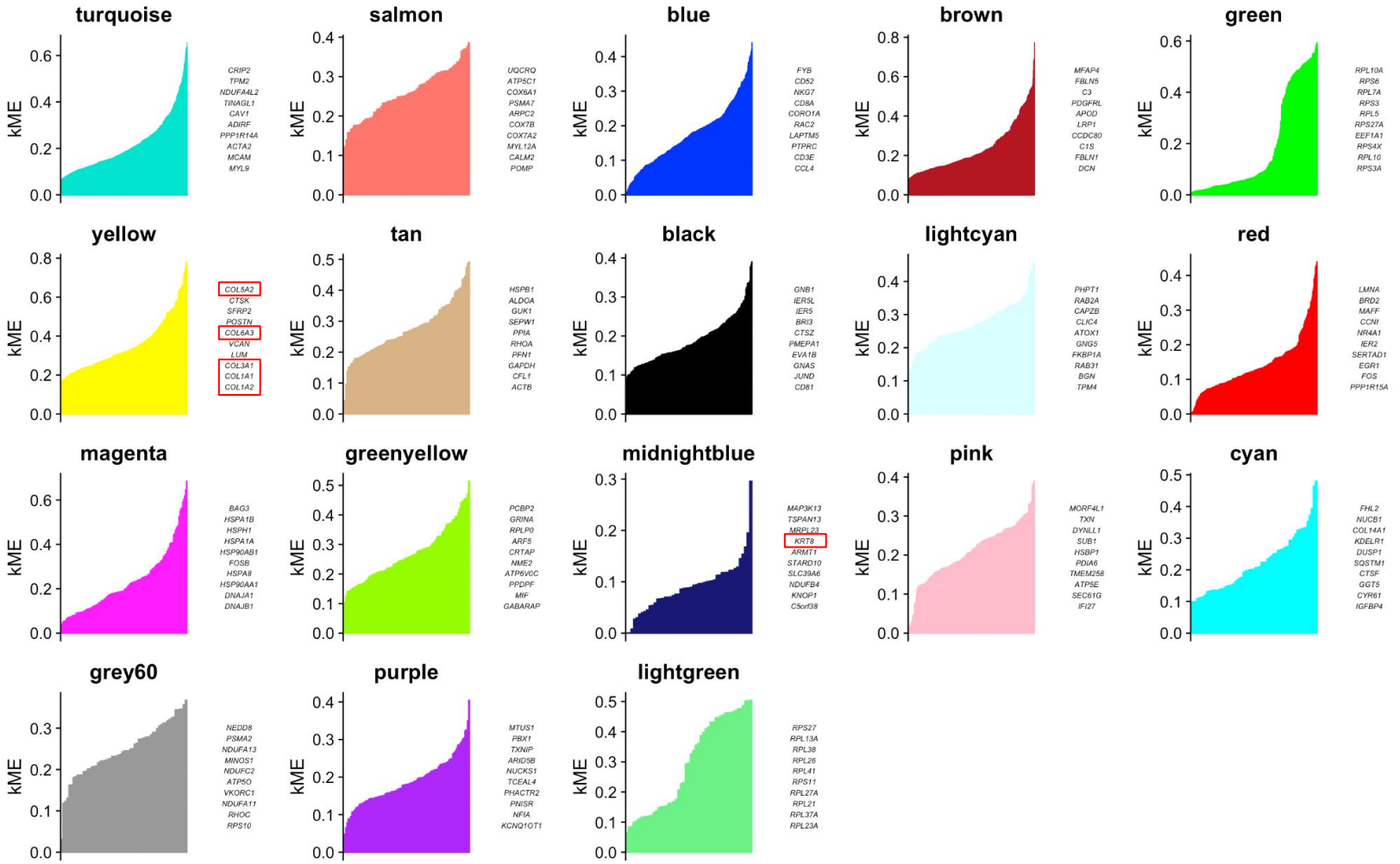
Co-expression Network Analysis

- hdWGCNA
  - Analysis tool for co-expression networks in high dimensional transcriptomics data
    - Co-expression networks
  - Approaches
    - kNN algorithm -> Metacells
    - Hierarchical clustering -> Dendrogram
      - Soft-power threshold
      - Modules
    - PCA -> Module eigengenes
    - Eigengene-based connectivity a.k.a kME
- EnrichR using MSigDb for hallmarks to understand our modules better

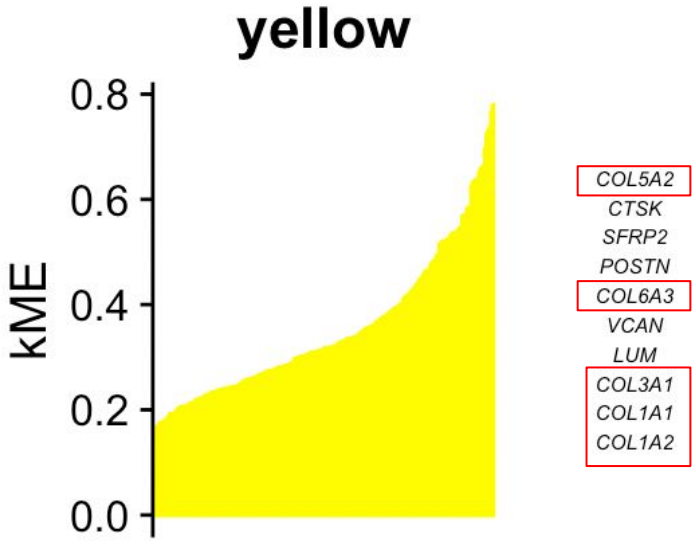
Soft-power threshold and Dendrogram

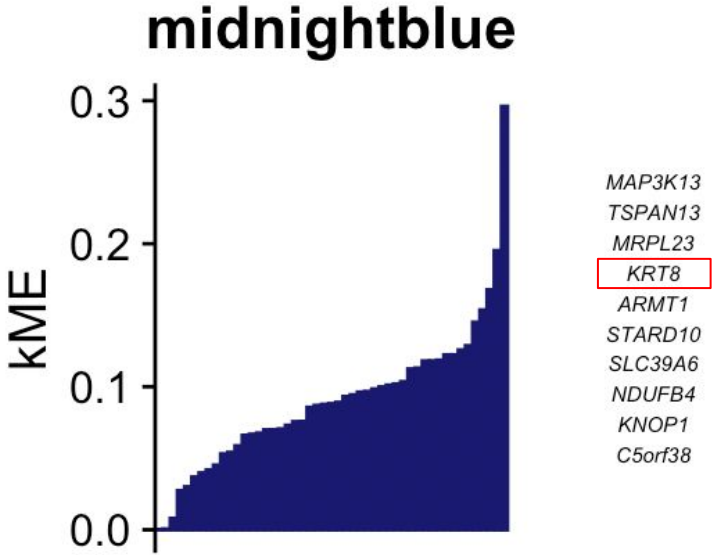





turquoise
kME
CRIP2
TPM2
NDUFA4L2
TINAGL1
CAV1
ADIRF
PPP1R14A
ACTA2
MCAM
MYL9
salmon
kME
UQCRQ
ATP5C1
COX6A1
PSMA7
ARPC2
COX7B
COX7A2
MYL12A
CALM2
POMP
blue
kME
FYB
CD52
NKG7
CD8A
CORO1A
RAC2
LAPTM5
PTPRC
CD3E
CCL4
brown
kME
MFAP4
FBLN5
C3
PDGFRL
APOD
LRP1
CCDC80
C1S
FBLN1
DCN
green
kME
RPL10A
RPS6
RPL7A
RPS3
RPL5
RPS27A
EEF1A1
RPS4X
RPL10
RPS3A
yellow
kME
COL5A2
CTSK
SFRP2
POSTN
COL6A3
VCAN
LUM
COL3A1
COL1A1
COL1A2
tan
kME
HSPB1
ALDOA
GUK1
SEPW1
PPIA
RHOA
PRN1
GAPDH
CFL1
ACTB
black
kME
GNB1
IER5L
IER5
BRI3
CTSZ
PMEPA1
EVA1B
GNAS
JUND
CD81
lightcyan
kME
PHPT1
RAB2A
CAPZB
CLIC4
ATOX1
GNG5
FKBP1A
RAB31
BGN
TPM4
red
kME
LMNA
BRD2
MAFF
CCNI
NR4A1
IER2
SERTAD1
EGR1
FOS
PPP1R15A
magenta
kME
BAG3
HSPA1B
HSPH1
HSPA1A
HSP90AB1
FOSB
HSPA8
HSP90AA1
DNAJA1
DNAJB1
greenyellow
kME
PCBP2
GRINA
RPLP0
ARF5
CRTAP
NME2
ATP6V0C
PPDPF
MIF
GABARAP
midnightblue
kME
MAP3K13
TSPAN13
MRPL23
KRT8
ARMT1
STARD10
SLC39A6
NDUFB4
KNOP1
C5orf38
pink
kME
MORF4L1
TXN
DYNLL1
SUB1
HSBP1
PDIA6
TMEM258
ATP5E
SEC61G
IFI27
cyan
kME
FHL2
NUCB1
COL14A1
KDELR1
DUSP1
SQSTM1
CTSF
GGT5
CYR61
IGFBP4
grey60
kME
NEDD8
PSMA2
NDUFA13
MINOS1
NDUFC2
ATP5O
VKORC1
NDUFA11
RHOC
RPS10
purple
kME
MTUS1
PBX1
TXNIP
ARID5B
NUCKS1
TCEAL4
PHACTR2
PNISR
NFIA
KCNQ1OT1
lightgreen
kME
RPS27
RPL13A
RPL38
RPL26
RPL41
RPS11
RPL27A
RPL21
RPL37A
RPL23A

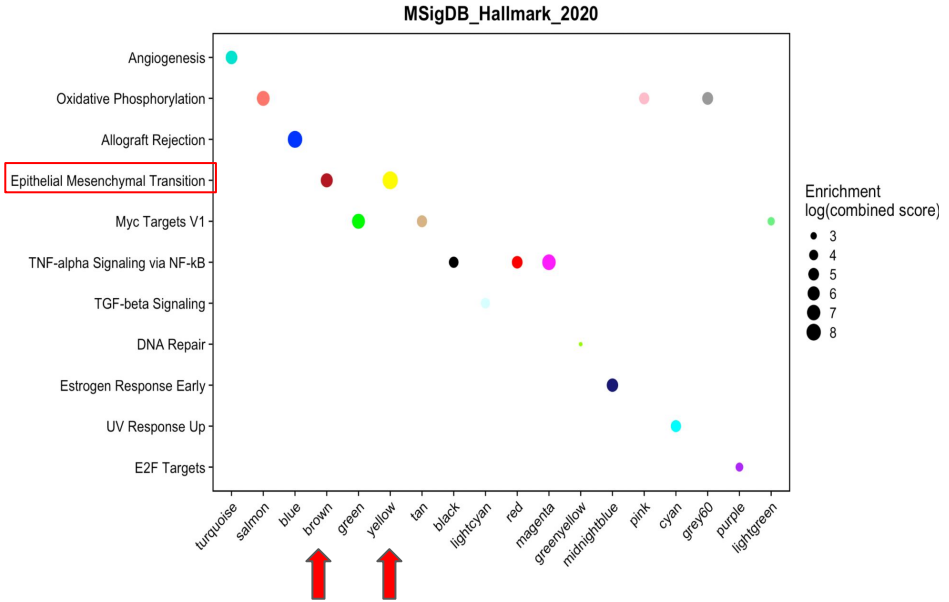
Enlarged plots with ranked genes by each module



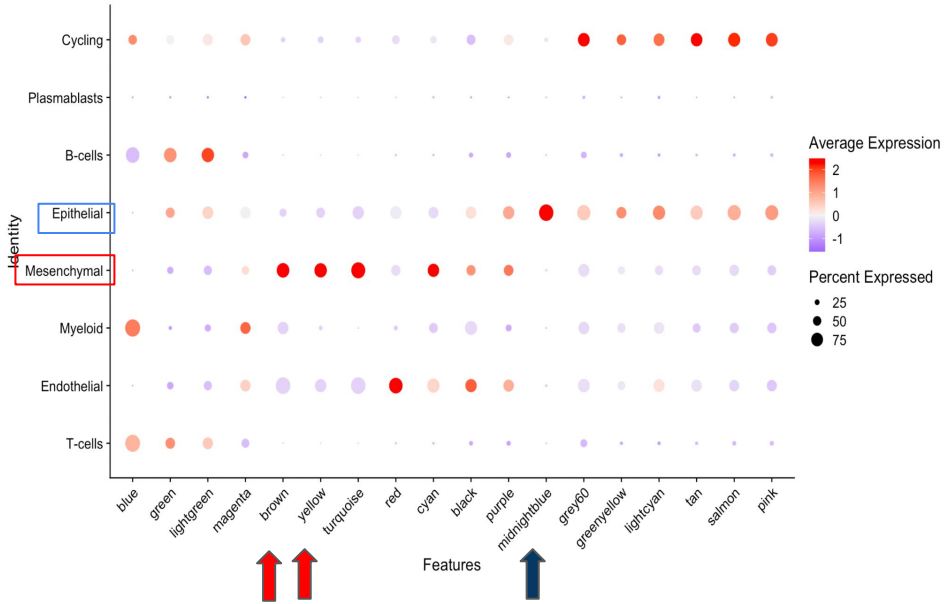


Dot plots for co-expressed gene modules

Across different hallmarks



Across different cell types



Discussion

- **Differential expression**
  - Angiogenesis markers have very low expression across all subtypes
  - Only CXCR4 chemokine is expressed in CXCR gene family
- **Co-expression Network**
  - The module having collagen type genes was highly expressed in the mesenchymal cells
  - The module was enriched in the epithelial mesenchymal transition which is often influenced by tumor microenvironment
- **Contradictory in EnrichR Barplot of differential expressed genes across tumor subtypes**
  - As the dataset was used for spatial transcriptomics, it might cause conflicting results.
  - It would be resolved if we further this analysis taking heterogeneity within the samples into account - divide dataset into specific dimensions or we incorporate the spatial analysis.

References

- Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J. R., Bartonicek, N., Wang, T., Larsson, L., Kaczorowski, D., Weisenfeld, N. I., Uytingco, C. R., Chew, J. G., Bent, Z. W., Chan, C. L., Gnanasambandapillai, V., Dutertre, C. A., … Swarbrick, A. (2021). A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, *53*(9), 1334–1347. https://doi.org/10.1038/s41588-021-00911-1
- Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., & Swarup, V. (2022). High dimensional co-expression networks enable discovery of transcriptomic drivers in complex biological systems. *bioRxiv*, 2022-09.
- Morabito, S., Miyoshi, E., Michael, N., Shahin, S., Martini, A. C., Head, E., ... & Swarup, V. (2021). Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nature genetics*, *53*(8), 1143-1155.

Supplementary Plots