SPATIAL TRANSCRIPTOMICS DATA ANALYSIS: THEORY AND PRACTICE

PRACTICAL SESSION 4

DR SIMON J COCKELL

ELEFTHERIOS (LEFTERIS) ZORMPAS

BIOSCIENCES INSTITUTE,

FACULTY OF MEDICAL SCIENCES,

NEWCASTLE UNIVERSITY

23/07/2023

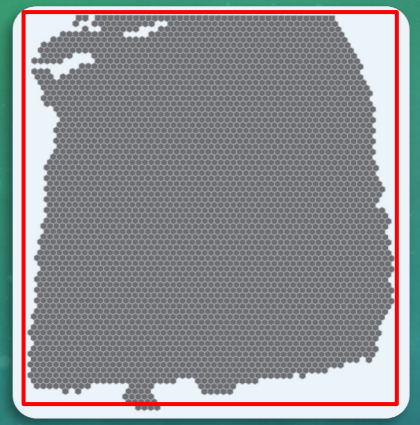
Practical session 4

In this practical session, we will have a hands-on exploration of GW-PCA and its application to STx data.

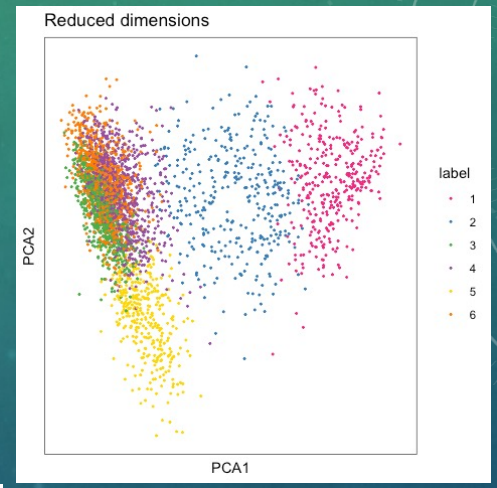
What can we learn from this novel technique?

4.1 GW-PCA: Geographically Weighted PCA

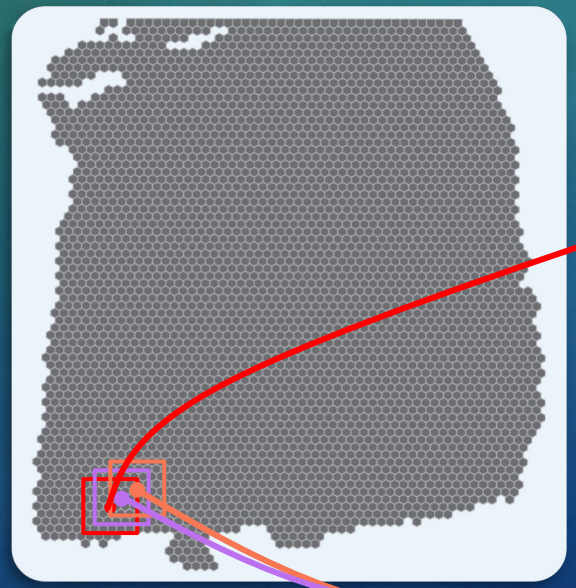
***GWPCA**: perform PCA analysis in a local way to reveal location-related principal components of variability*

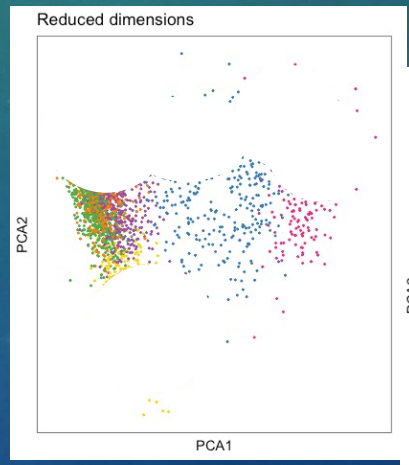


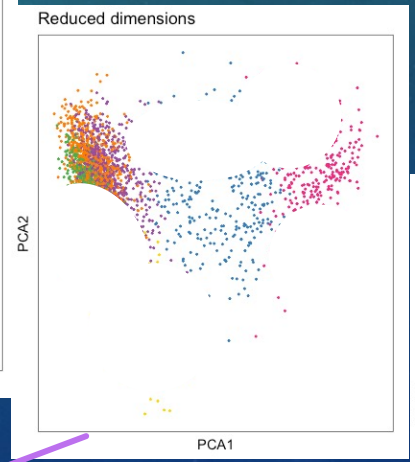
GLOBAL
–CLASSIC–
PCA

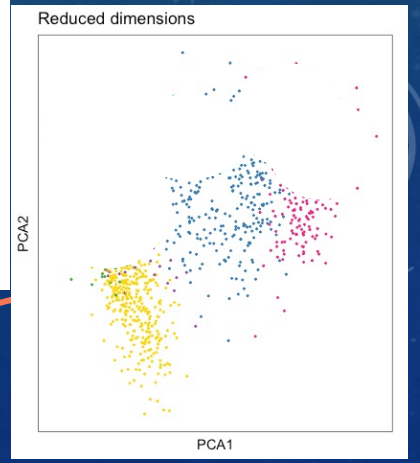


GWPCA: “moving windows”









4.2 Load Quality Controlled and Normalised data

```
sfe <- readRDS(file = "./data/to_load/practical03_sfe.rds")
top_hvgs <- readRDS(file = "./data/to_load/practical03_topHVGs.rds")
```

4.4 Parameter preparation for GWPCA

```
## Get the gene names that are going to be evaluated
vars = top_hvgs
## Set a fixed bandwidth
bw = 6*sfe@metadata[["spotDiameter"]][["JBO019"]][["spot_diameter_fullres"]]
## Set the number of components to be retained
k = 20
## Set the kernel to be used
kernel = "gaussian"
## Set the Minkowski distance power: p = 2 --> Euclidean
p = 2
## Is the bandwidth adaptive?: No because spots are fixed
adaptive = FALSE
## Cross-Validate GWPCA?
cv = TRUE
## Calculate PCA scores?
scores = FALSE
## Run a robust GWPCA?
robust = FALSE
## Make a cluster for parallel computing (otherwise GWPCA is slow!)
my.cl <- parallel::makeCluster(parallelly::availableCores() - 1, type = 'FORK')
```

4.5 Run GWPCA

```
# DO NOT RUN THIS CHUNK
pcagw <- gwpcaSTE(sfe = sfe,
                  assay = "logcounts",
                  vars = vars,
                  p = p,
                  k = k,
                  bw = bw,
                  kernel = kernel,
                  adaptive = adaptive,
                  scores = scores,
                  robust = robust,
                  cv = cv,
                  future = FALSE,
                  strategy = "cluster",
                  workers = my.cl,
                  verbose = FALSE)
```

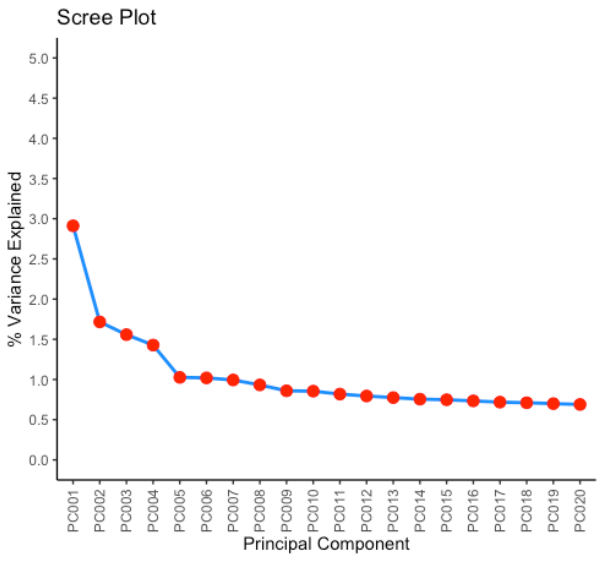
Because GWPCA can take some time to run, we ran it for you and below you can load the output:

```
pcagw <- readRDS(file = "./data/to_load/practical04_pcagw.rds")
```

gwpcaSTE:

- Function from the `STExplorerDev` package.
- Re-implementation of the `gwpca` function from the `GWmodel` package.
- Sets a `future` backend to allow parallel processing.
- The `future`, `strategy` and `workers` arguments are used to set up the parallel backend.
- By default runs sequentially.

4.6 Plot global PCA results



The percentages of variance explained by the global PCA PCs are small.

- If the first 4 PCs explain less than 15% of the variance then:
  - the data is highly dispersed or
  - there is a large amount of noise or
  - lack of clear structure in the data or
  - lack of meaningful patterns

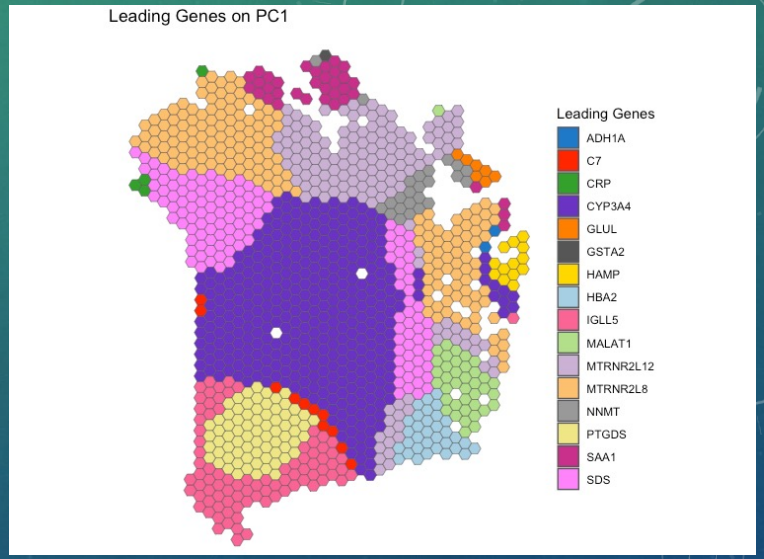
GWPCA might be more appropriate because the global model might not reflect what is happening locally

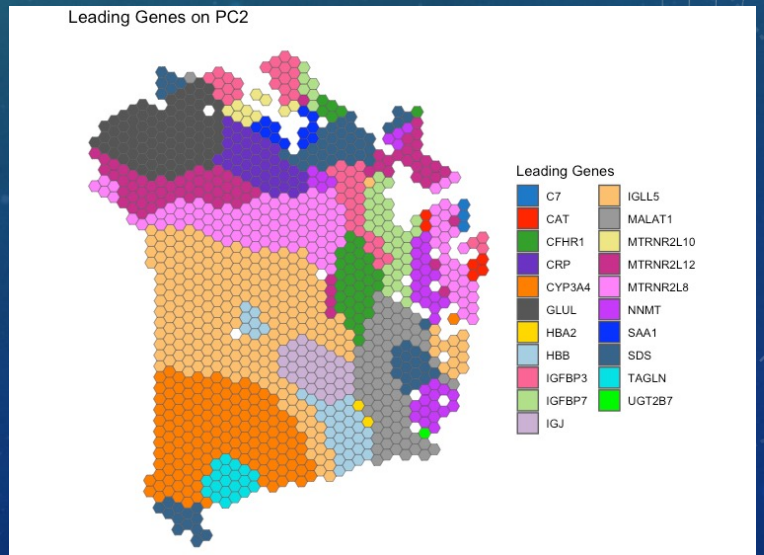
4.7 Identify the leading genes in each location

Single leading gene

```
## Extract leading genes
pcagw <- gwpca_LeadingGene(gwpca = pcagw,
                           sfe = sfe,
                           pc_nos = 1:4,
                           type = "single",
                           names = "gene_names")
```

```
## 16  leading genes found for  PC1
## The leading genes in  PC1  are:
##     ADH1A        C7       CRP    CYP3A4      GLUL     GSTA2      HAMP      HBA2
##         2        11         4       365         7         1        13        33
##     IGLL5    MALAT1 MTRNR2L12  MTRNR2L8     NNMT     PTGDS      SAA1       SDS
##        87        39       153       181        23        73        36       133
## 21  leading genes found for  PC2
## The leading genes in  PC2  are:
##        C7       CAT     CFHR1       CRP    CYP3A4      GLUL      HBA2       HBB
##         3         6        38        39       149        83         2        37
##    IGFBP3    IGFBP7       IGJ     IGLL5    MALAT1 MTRNR2L10 MTRNR2L12  MTRNR2L8
##        49        39        34       246        80        10        78       124
##      NNMT      SAA1       SDS     TAGLN    UGT2B7
##        42        12        69        20         1
## 24  leading genes found for  PC3
## The leading genes in  PC3  are:
##     AEBP1        C7       CAT     CFHR1       CRP    CYP3A4      GLUL      HBA2
##         2         2        27        20         5        20        17        27
##       HBB    IGFBP3    IGFBP7       IGJ     IGLL5    MALAT1 MTRNR2L10 MTRNR2L12
##       150        41        77         6       399       136         6        61
##  MTRNR2L8      MYL9      NNMT      SAA1   SCGB3A1       SDS     TAGLN    UGT2B7
##        25         9        24         6        56        15        26         4
## 25  leading genes found for  PC4
## The leading genes in  PC4  are:
##     AEBP1       CAT     CFHR1       CRP      FXYD2      GLUL     GSTA2      HBA2
##         1        53        15         7         7        33         3         2
##       HBB    IGFBP3    IGFBP7       IGJ     IGLL5    MALAT1 MTRNR2L10 MTRNR2L12
##       181       100        51        60       281       201         5        16
##  MTRNR2L8      MYL9      NNMT      ORM2      SAA1       SDS    SPINK1     TAGLN
##        16         5        55         6         6        37        12         4
##    UGT2B7
##         4
```





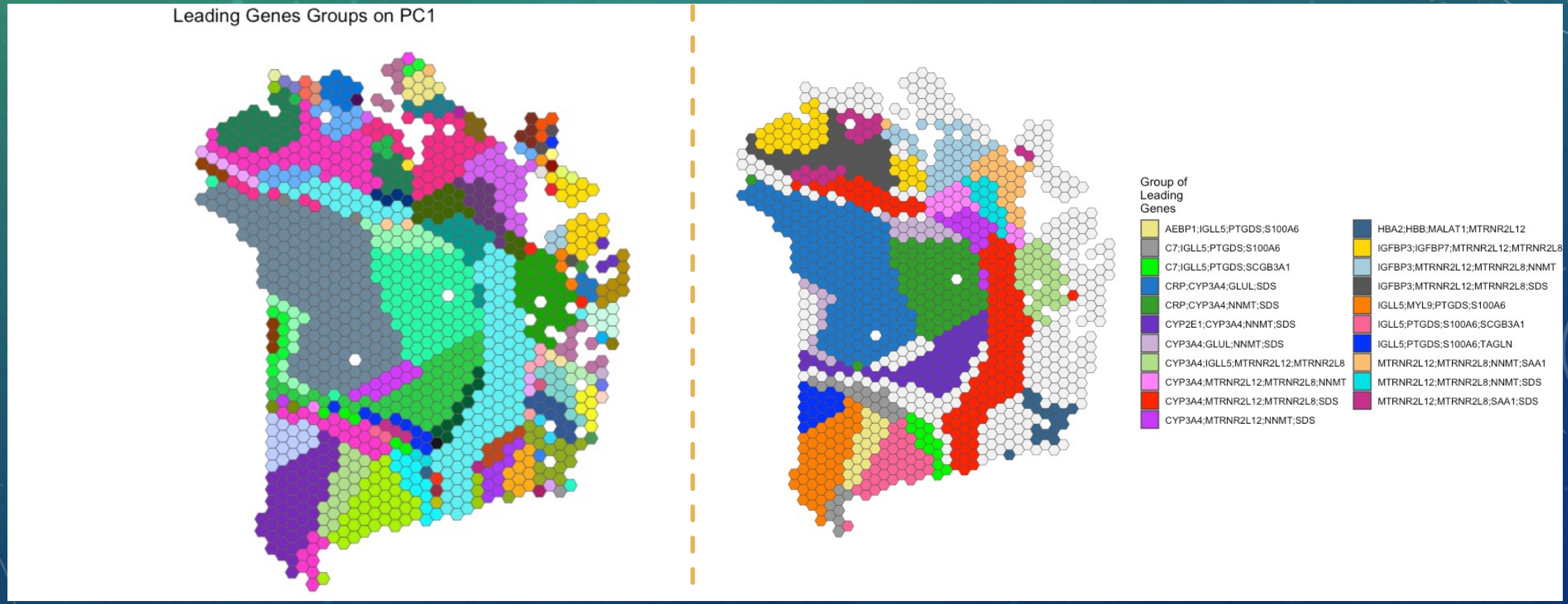
4.7 Identify the leading genes in each location

Top-k leading genes

```
pcagw <- gwpca_LeadingGene(gwpca = pcagw,
                           sfe = sfe,
                           pc_nos = 1:4,
                           genes_n = 4,
                           type = "multi",
                           method = "membership",
                           names = "gene_names")
```

```
## The number of individual leading genes groups found for PC1 is: 110
## These groups are: Too many to print them!
## The number of individual leading genes groups found for PC2 is: 240
## These groups are: Too many to print them!
## The number of individual leading genes groups found for PC3 is: 310
## These groups are: Too many to print them!
## The number of individual leading genes groups found for PC4 is: 421
## These groups are: Too many to print them!
```

Too many groups to print them out as we did at the previous ones

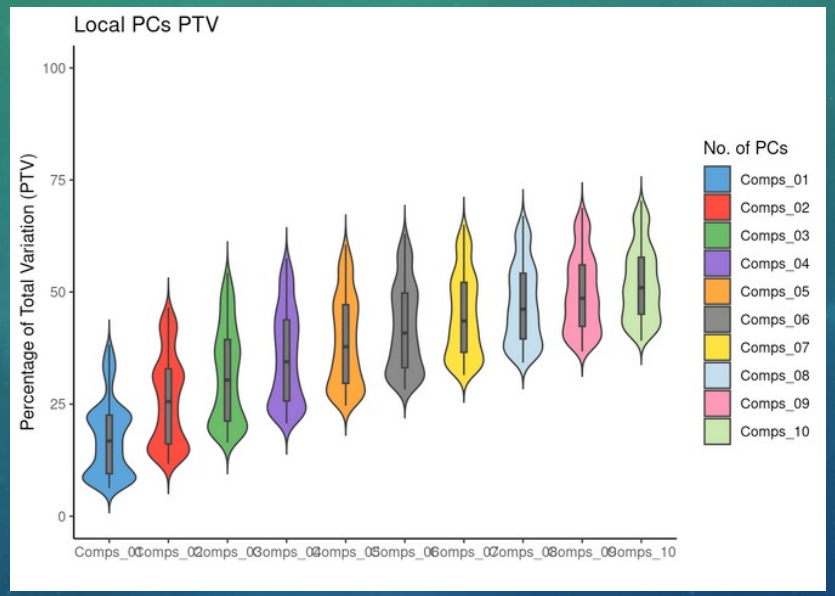


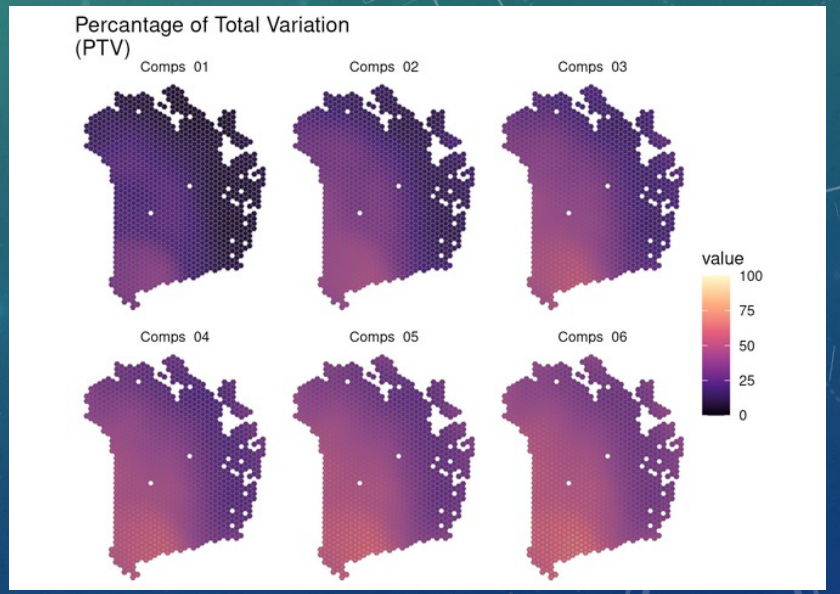
4.8 Percentage of Total Variation (PTV)

```
## Calculate the PTV for multiple Components
pcagw <- gwpca_PropVar(gwpca = pcagw, n_comp = 2:10, sfe = sfe)
```

```
##     Comps_01         Comps_02         Comps_03         Comps_04
##  Min.   : 6.279   Min.   :11.67   Min.   :16.43   Min.   :20.69
##  1st Qu.: 9.483   1st Qu.:16.13   1st Qu.:21.24   1st Qu.:25.69
##  Median :16.782   Median :25.54   Median :30.37   Median :34.46
##  Mean   :17.370   Mean   :25.92   Mean   :31.35   Mean   :35.49
##  3rd Qu.:22.534   3rd Qu.:32.87   3rd Qu.:39.42   3rd Qu.:43.81
##  Max.   :38.254   Max.   :46.50   Max.   :54.25   Max.   :57.51
##     Comps_05         Comps_06         Comps_07         Comps_08
##  Min.   :24.64   Min.   :28.28   Min.   :31.49   Min.   :34.26
##  1st Qu.:29.65   1st Qu.:33.13   1st Qu.:36.54   1st Qu.:39.53
##  Median :37.79   Median :40.86   Median :43.53   Median :46.17
##  Mean   :38.98   Mean   :42.07   Mean   :44.84   Mean   :47.38
##  3rd Qu.:47.17   3rd Qu.:49.78   3rd Qu.:52.16   3rd Qu.:54.19
##  Max.   :60.60   Max.   :62.97   Max.   :65.04   Max.   :67.03
##     Comps_09         Comps_10
##  Min.   :36.76   Min.   :39.15
##  1st Qu.:42.34   1st Qu.:45.05
##  Median :48.60   Median :50.96
##  Mean   :49.73   Mean   :51.91
##  3rd Qu.:56.07   3rd Qu.:57.77
##  Max.   :68.83   Max.   :70.39
```

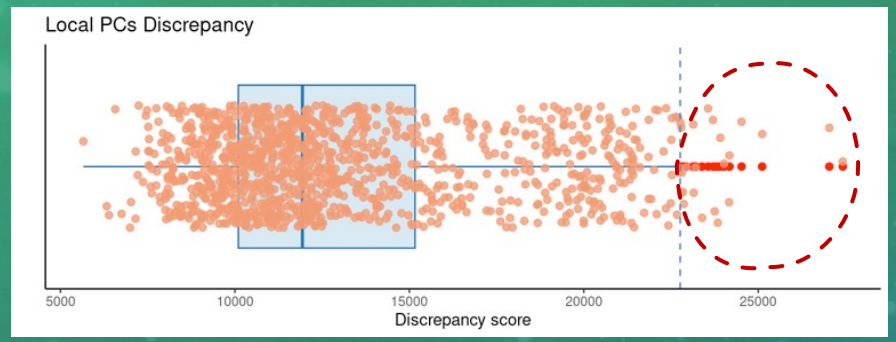
Remember these are cumulative %...

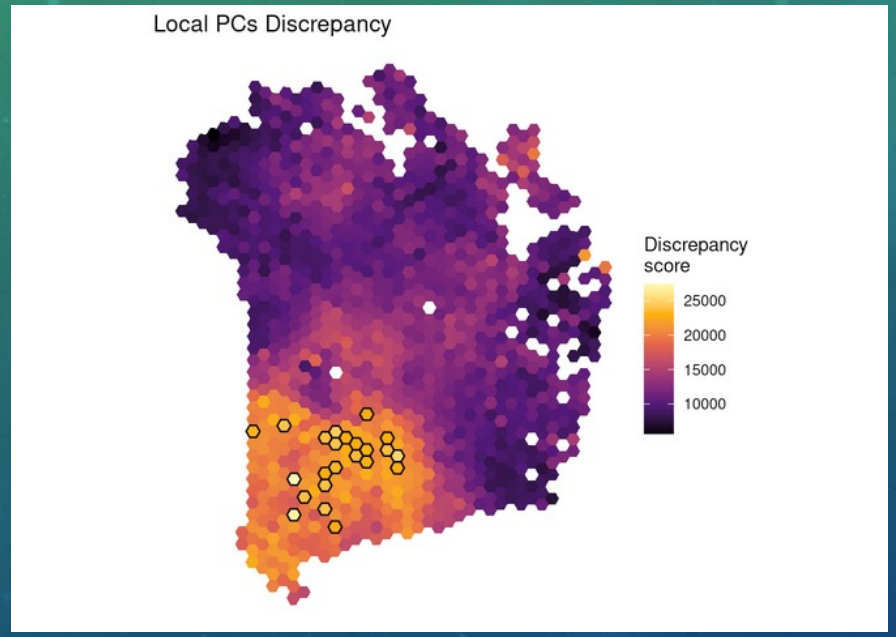


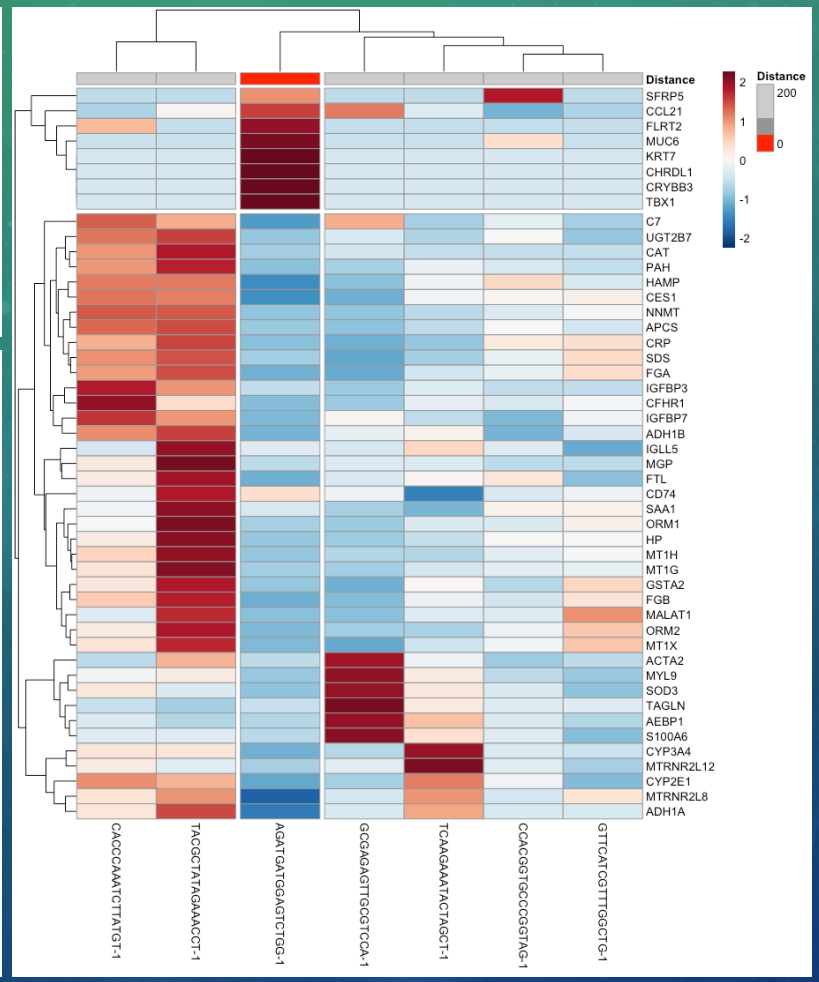


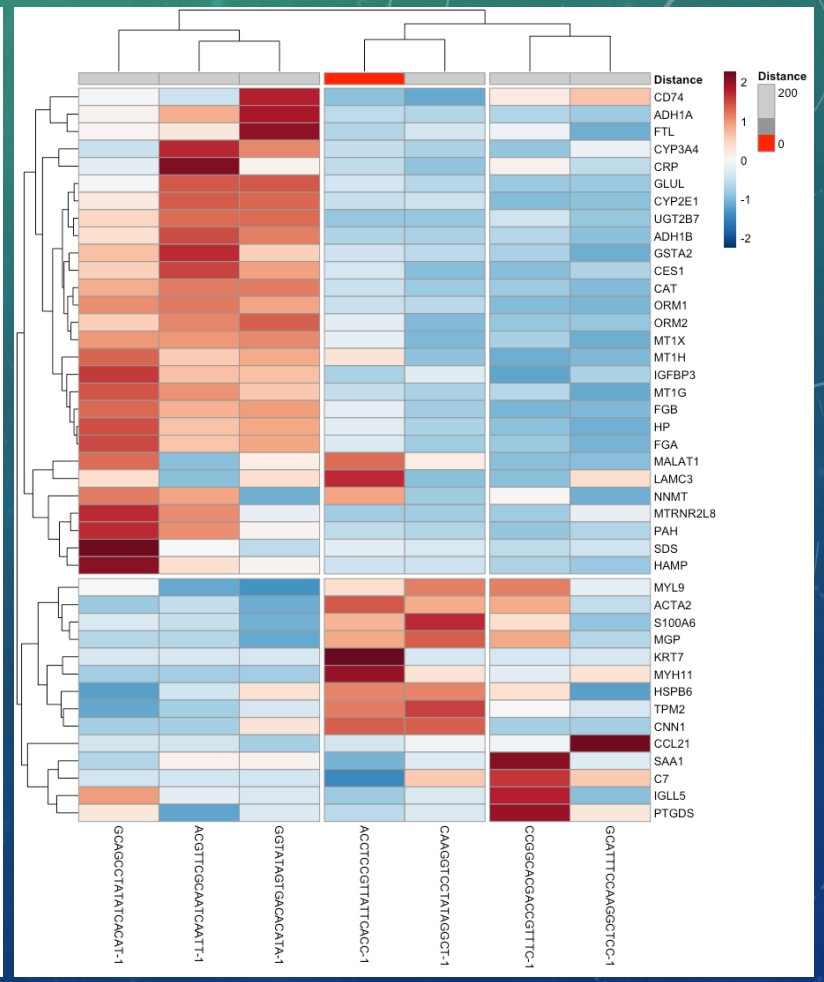
4.9 Identify discrepancies

```
## Plot the discrepancies as boxplot
plotGWPCA_discr(pcagw, type = "box")
```









AKNOWLEDGEMENTS

Eleftherios Zormpas

Newcastle University

Dr Simon J Cockell

Newcastle University

Dr Rachel Queen

Newcastle University

Prof. Alex Comber

UNIVERSITY OF LEEDS

iSMB feedback form:

ISMB ECCB 2023 LYON FRANCE

Newcastle University

Medical Research Council

MRC DiMeN Doctoral Training Partnership