

# Intégration de données



Vincent Guillemot

Cathy Philippe

# Objectif général

- .Comprendre les **enjeux méthodologiques et pratiques** de l'intégration de données multi-omiques
- .Utiliser ces connaissances pour **concevoir une stratégie d'analyse intégrative efficace** dans leurs propres projets de recherche.

# Objectifs pédagogiques spécifiques

## Objectifs liés aux méthodes d'intégration en général

- .Cartographier les approches classiques
- .Connaître les résultats que l'on obtient en sortie d'une ACP

## Objectifs liés à RGCCA

- .Paramétrer RGCCA en fonction du plan d'expérience.
- .Comprendre les paramètres de base de RGCCA
- .Connaître et savoir décrypter les résultats obtenus par RGCCA
- .Connaître et savoir décrypter les résultats obtenus par SGCCA

Quelles sont les différentes questions que vous souhaitez aborder avec des expériences (multi-)omiques ?

Notez vos réponses dans le tableau du script partagé

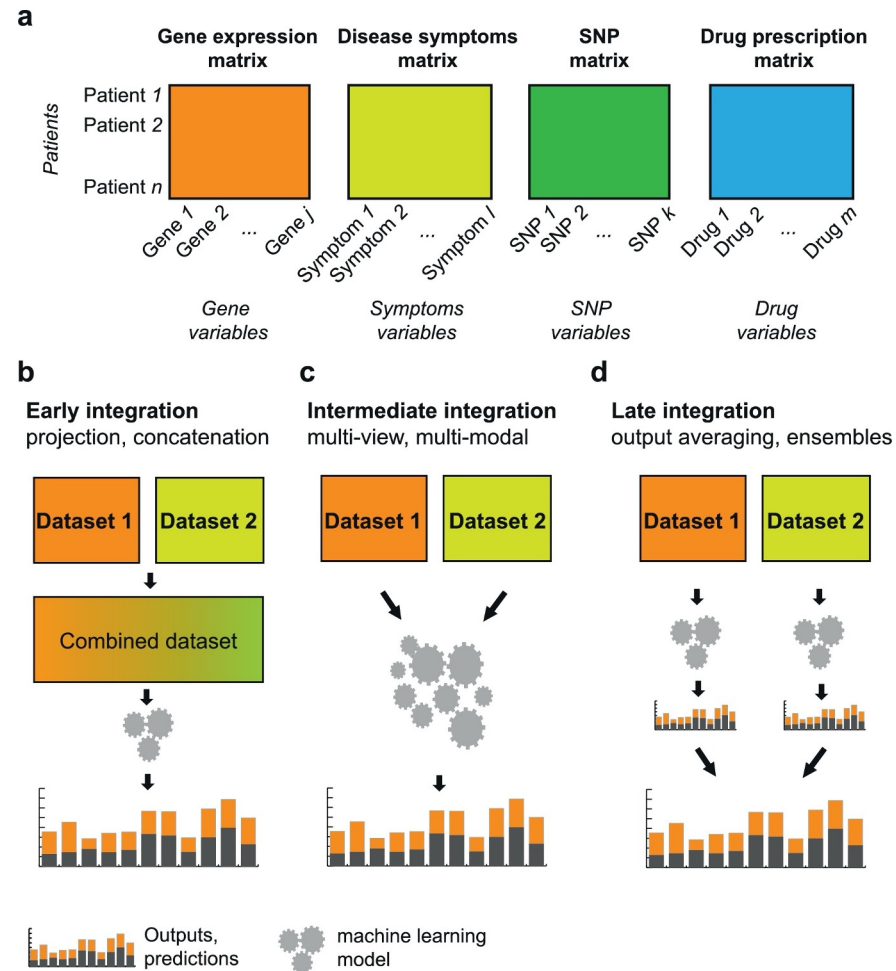
[https://docs.google.com/document/d/13deR-EVoFd\\_siSKun8hzbwaG7ywJAe5wmcFjP2ZS1iM/edit?tab=t.0](https://docs.google.com/document/d/13deR-EVoFd_siSKun8hzbwaG7ywJAe5wmcFjP2ZS1iM/edit?tab=t.0)

# Activité Wooclap ! (youpi)

<https://app.wooclap.com/FMEBBT?from=event-page>

Quelles sont les méthodes factorielle et/ou d'intégrations de données que vous connaissez/utilisez ?

# Différentes manières d'intégrer



# Avantages et inconvénients

| Type d'intégration | Avantages                                                                         | Inconvénients                                          |
|--------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------|
| Early              | .considère toutes interactions possibles entre toutes les variables               | .a tendance à négliger les blocs avec peu de variables |
| Intermediate       | .préserve la structure des données,<br>.traite chaque bloc avec ses spécificités. | .nécessite un développement méthodologique ad hoc      |
| Late               | .traitement spécifique à chaque modalité                                          | .2 étapes d'analyse → accumulation des erreurs         |

# Méthodes d'analyses

- .Carte mentale

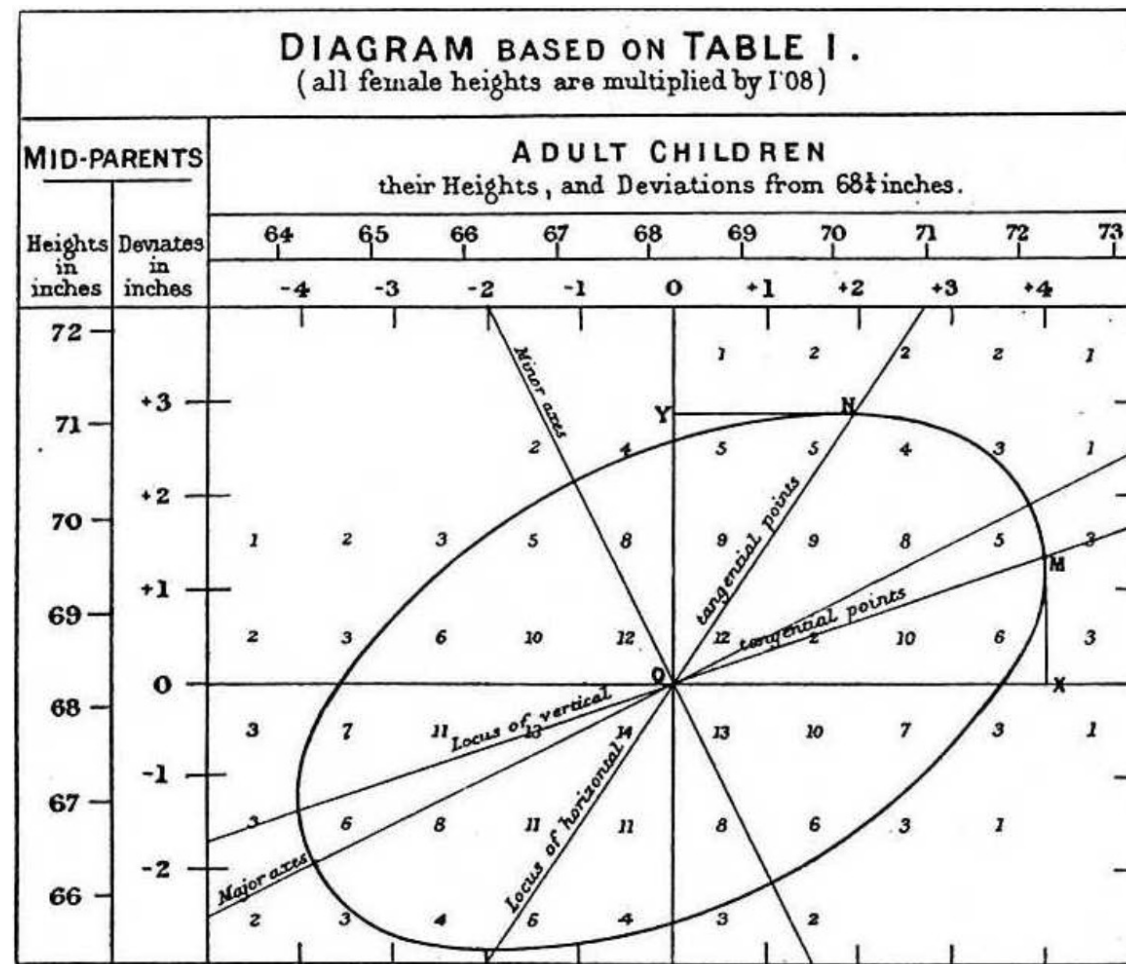
- .<https://framindmap.org/c/maps/1519259/public>

- .Ou cliquez sur le lien du script partagé

- .Tableau 3ème colonne : quelle analyse pour ma question ?

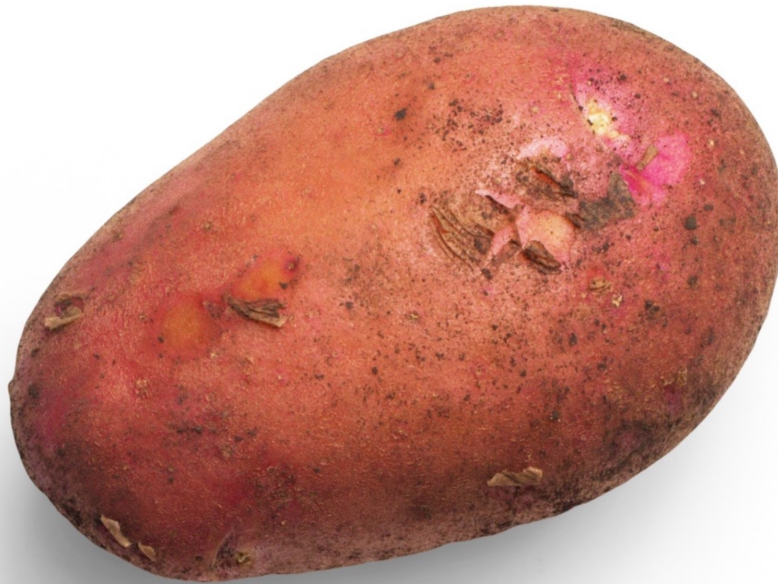
- .Tableau 4ème colonne : early, intermediate ou late ?

# Analyse en Composantes Principales



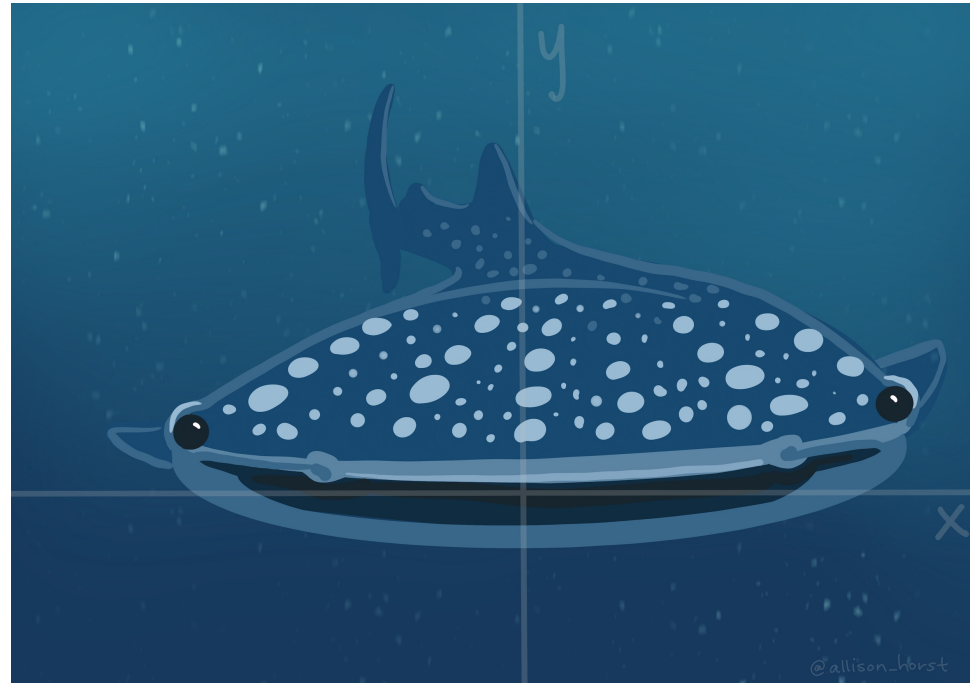
from Galton 1859

# Analyse Coupe-Patate



Cut the yummiest French fries

# Requin baleine (vous)...



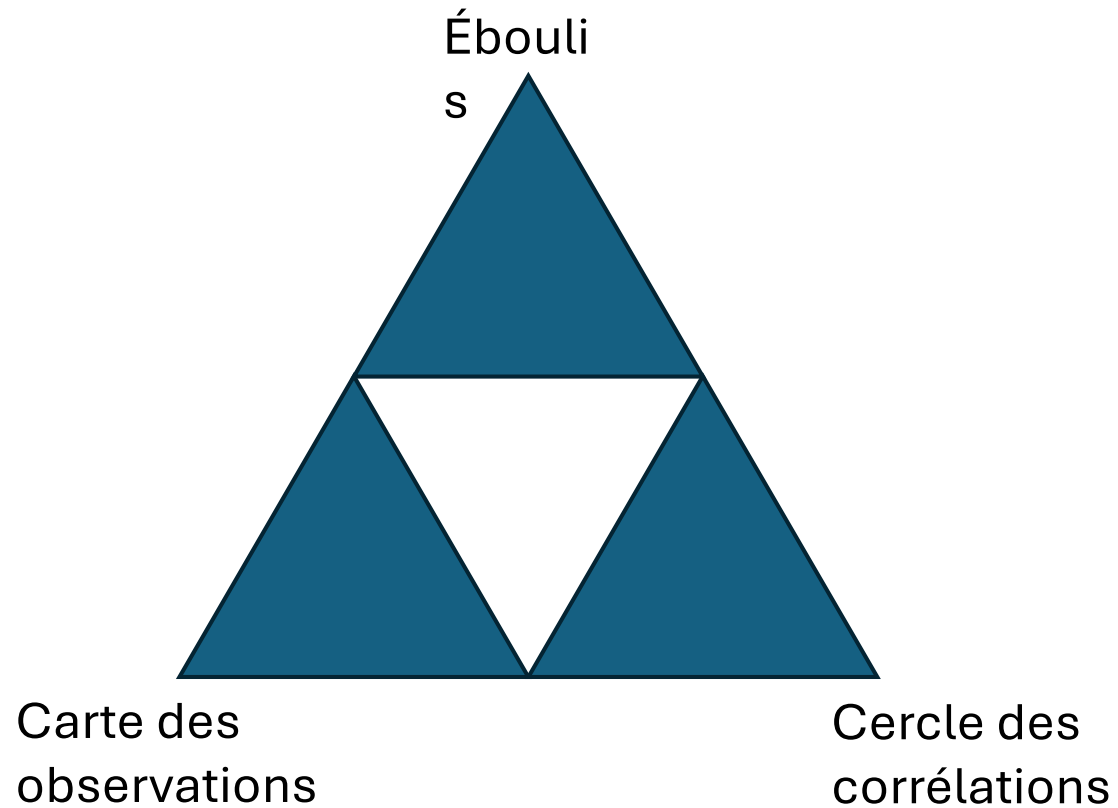
J'ai repéré des crevettes!

# Requin baleine (vous)... versus krill (data)



Je vais en manger le plus!

# La tri-force de l'ACP



# Français et anglais

“Aaaaah, mais acépé en fait c’est la PCA !”

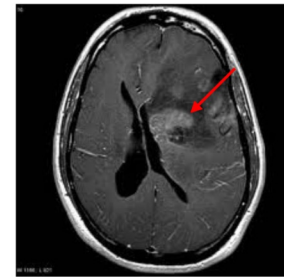
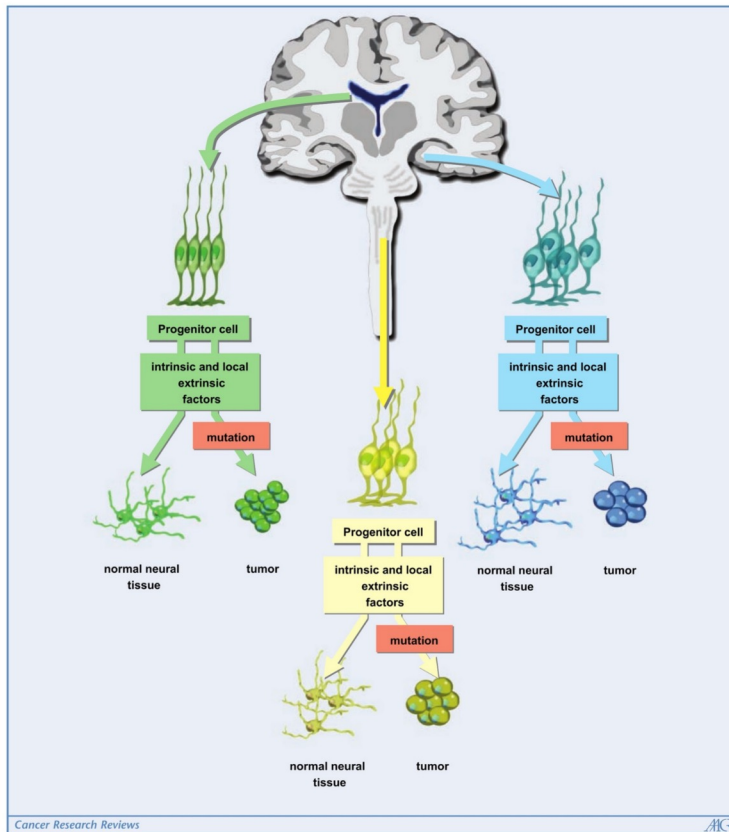
*(Anonyme, EBAll niveau 2, 2023)*

| English                              | Français                                   |
|--------------------------------------|--------------------------------------------|
| PCA = principal component analysis   | ACP = analyse en composantes principales   |
| SVD = singular value decomposition   | SVD = décomposition en valeurs singulières |
| EVD = eigenvalue decomposition       | décomposition en éléments propres          |
| ICA = independent component analysis | ICA = analyse en composantes indépendantes |
| MDS = multidimensional scaling       | MDS = multidimensional scaling             |

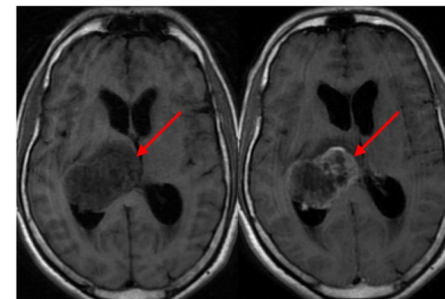
Les données  
d'aujourd'hui

# Localisation des gliomes malins pédiatriques

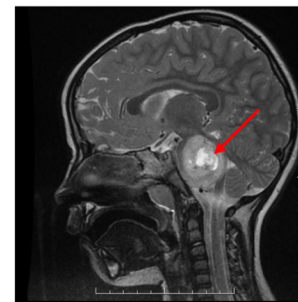
Gilbertson et al. (2007)



Hémisphère



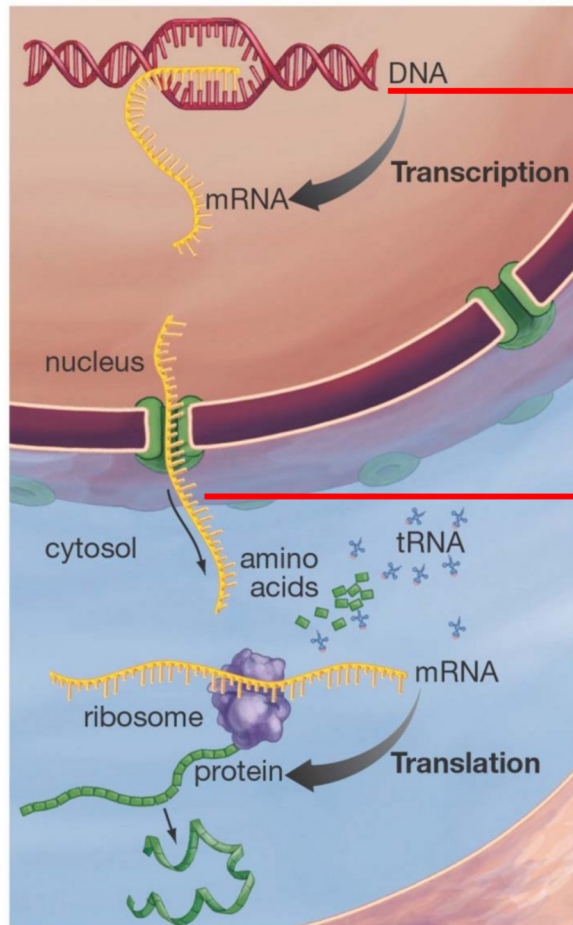
Ligne médiane



Tronc cérébral

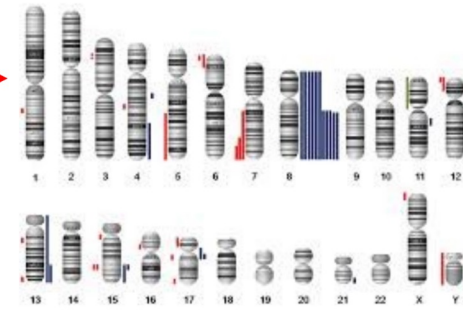
# Les données omiques disponibles

<https://sites.google.com/site/bentzingerbiology/biochemistry/8-dna-to-proteins>



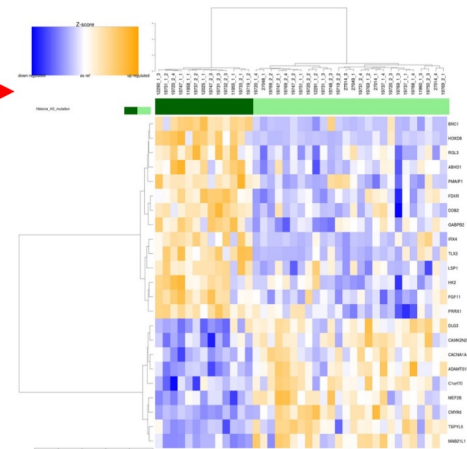
Aberrations  
du nombre  
de copies

Comparative  
Genomique  
Hybridization



Expression  
des gènes

Gene Expression



# Questions biologiques

## **.Emplacement de la tumeur (hémisphères, ligne médiane, tronc cérébral)**

.Quels sont les gènes et les segments caractéristiques de chaque localisation, qui pourraient nous renseigner sur la cellule d'origine des tumeurs ?

## **.Survie globale**

.L'expression des gènes et les aberrations du nombre de copies ont-elles une influence sur la survie globale ? si oui, laquelle ?

Améliorer la classification des gliomes malins pédiatriques, notamment son pouvoir pronostique.

# Les données gliomaData

```
## Changer le répertoire de travail  
setwd("LE_CHEMIN")
```

```
## Copie des données  
file.copy(  
  from = "/shared/projects/2514_ebiii_n2/integration",  
  to = ".",  
  recursive = TRUE)
```

```
### Puis aller à → scripts/data_01_lecture.R
```

# Analyse en composantes principales de l'expression des gènes

## La fonction `prcomp` de R-base

```
res.pca <- prcomp(gex, scale. = FALSE)
names(res.pca)
[1] "sdev"      "rotation" "center"    "scale"
"x"
```

# Activité : association

<https://app.wooclap.com/FMEBBT?from=event-page>

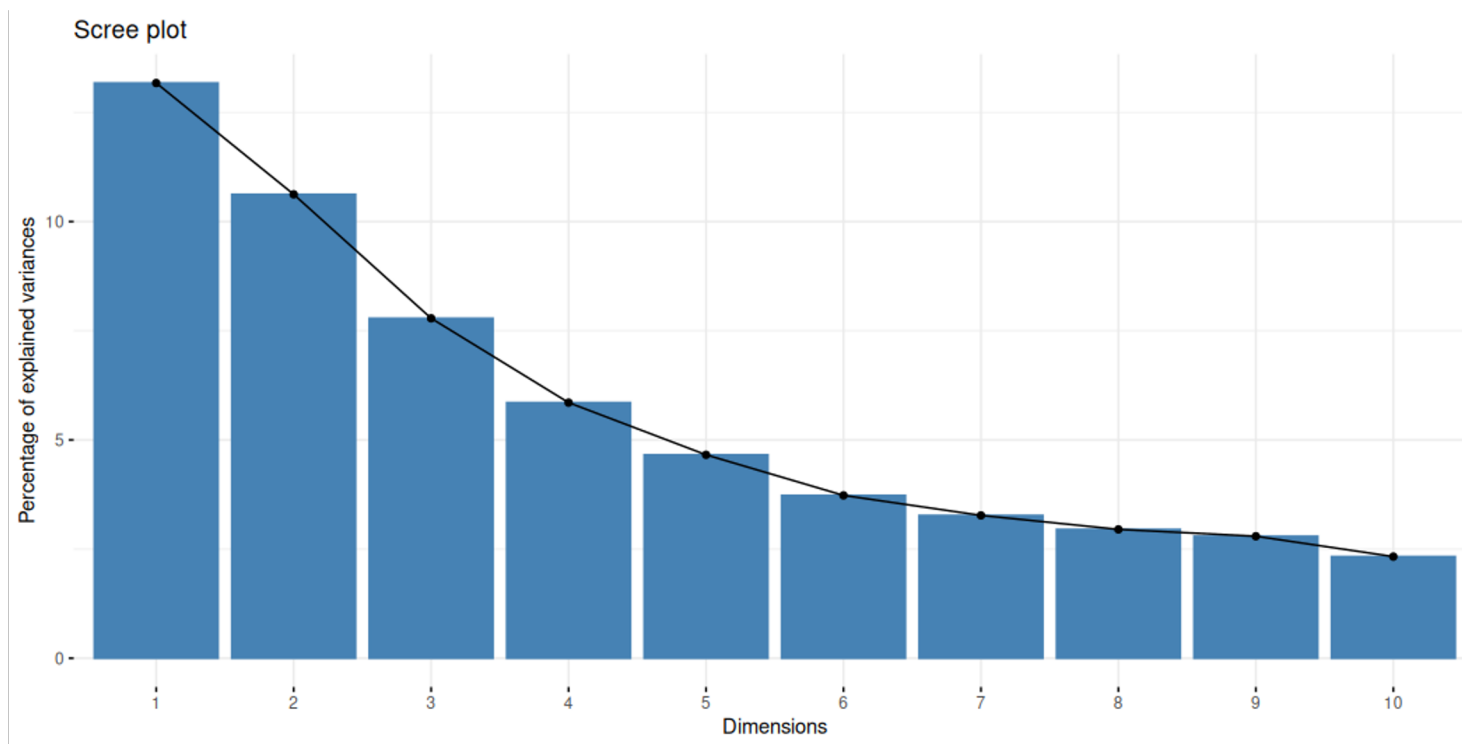
A quelles sorties de l'Analyse en Composantes Principales correspondent ces graphes ?

# Visualisation de l'ACP

```
library(factoextra)
```

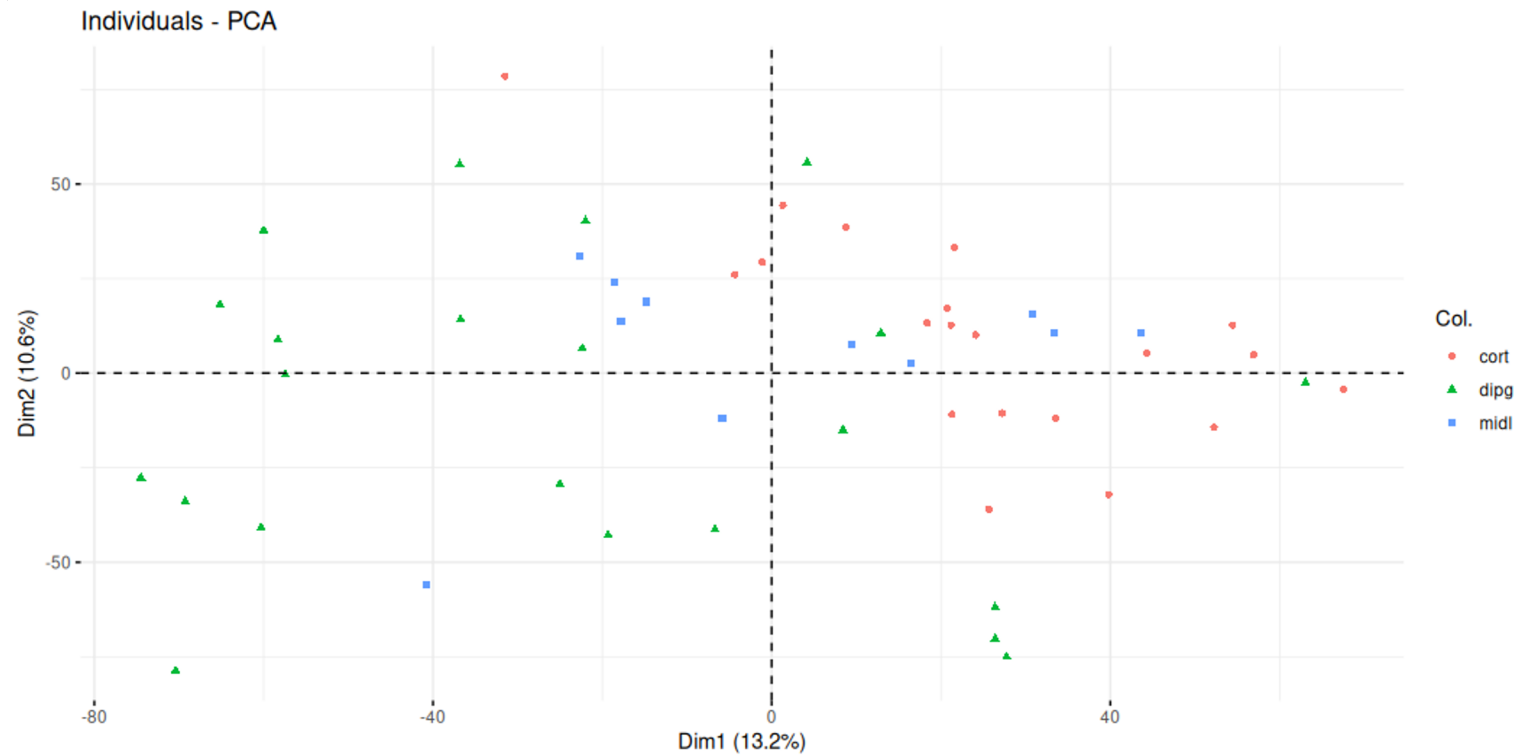
# Le screeplot

```
fviz_screepLOT(res.pca)
```



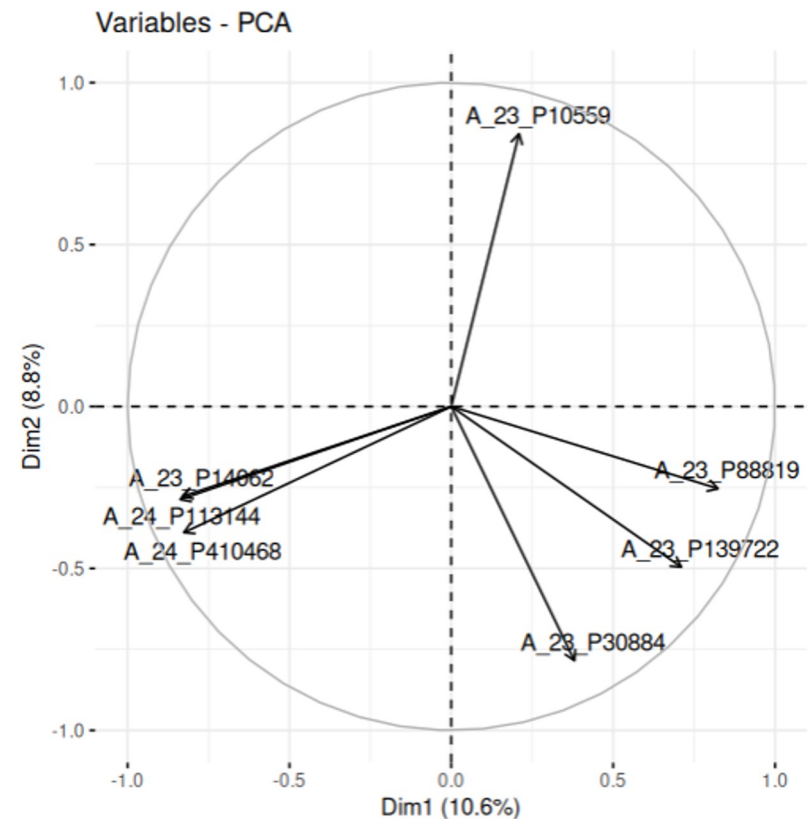
# Les observations

```
fviz_pca_ind(res.pca, geom = "point", col.ind = loc, invisible = "quali")
```



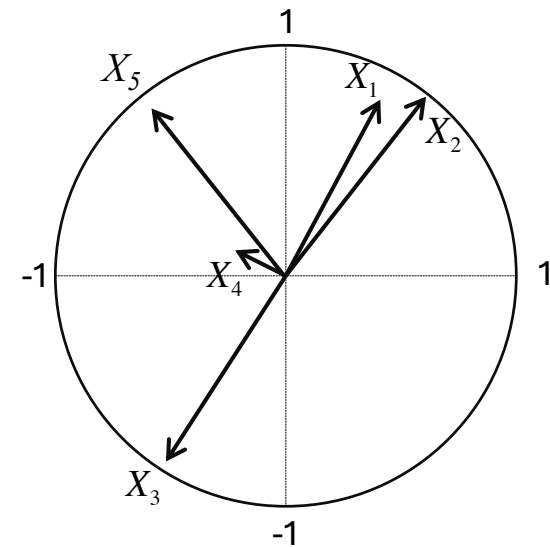
# Le cercle des corrélations (variables)

```
res.pca.scale <- prcomp(gex,  
scale. = TRUE)  
  
fviz_pca_var(res.pca.scale,  
select.var = list(contrib = 7),  
repel = TRUE)
```



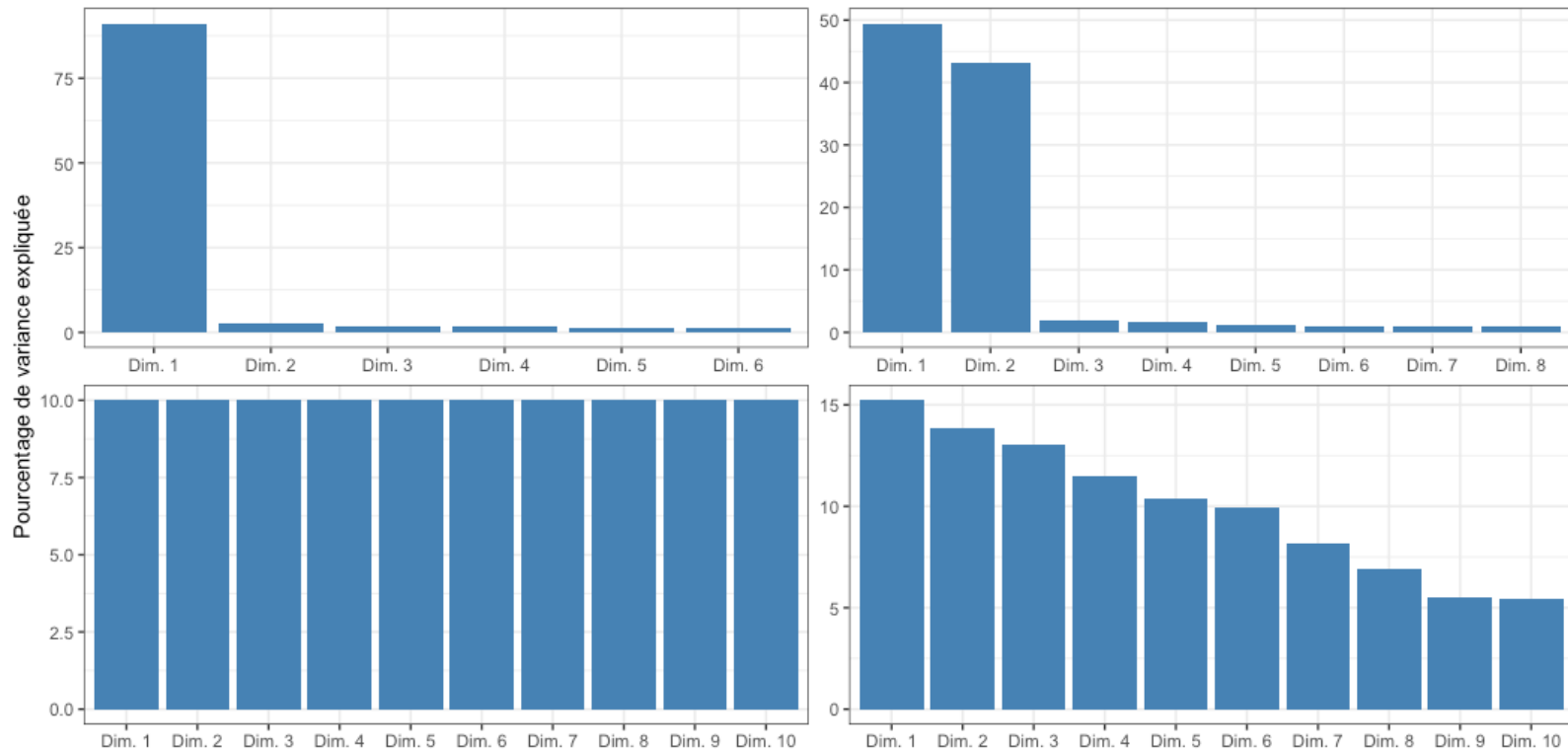
## Interpretation of the PCA plot of the individuals using the variables

- An arrow close to the circle indicates a variable that is well characterized by the represented PCs ( $X_1, X_2, X_3, X_5$ ).
- An arrow close to the center indicates a variable whose properties are not captured/described by the represented PCs ( $X_4$ ).
- Two arrows close to the circle and to each other indicate a positive correlation between the two variables ( $X_1, X_2$ ).
- Two arrows close to the circle and in opposite direction indicates a negative correlation between the two variables ( $X_1, X_3$  or  $X_2, X_3$ ).
- Two arrows close to the circle and at a right angle to one another are not correlated ( $X_1$  with  $X_5$  or  $X_2$  with  $X_5$ ).

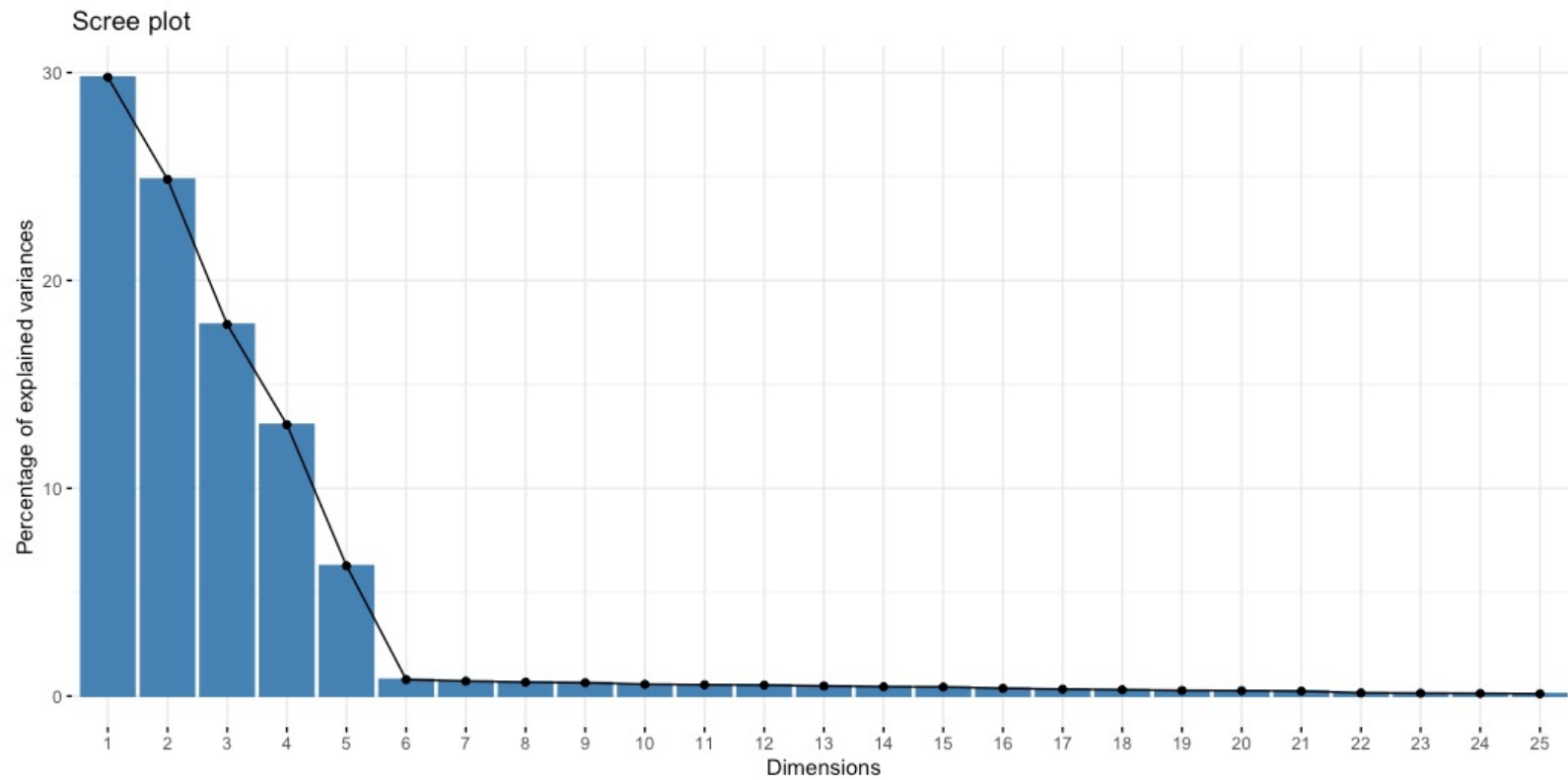


# Les alertes graphiques

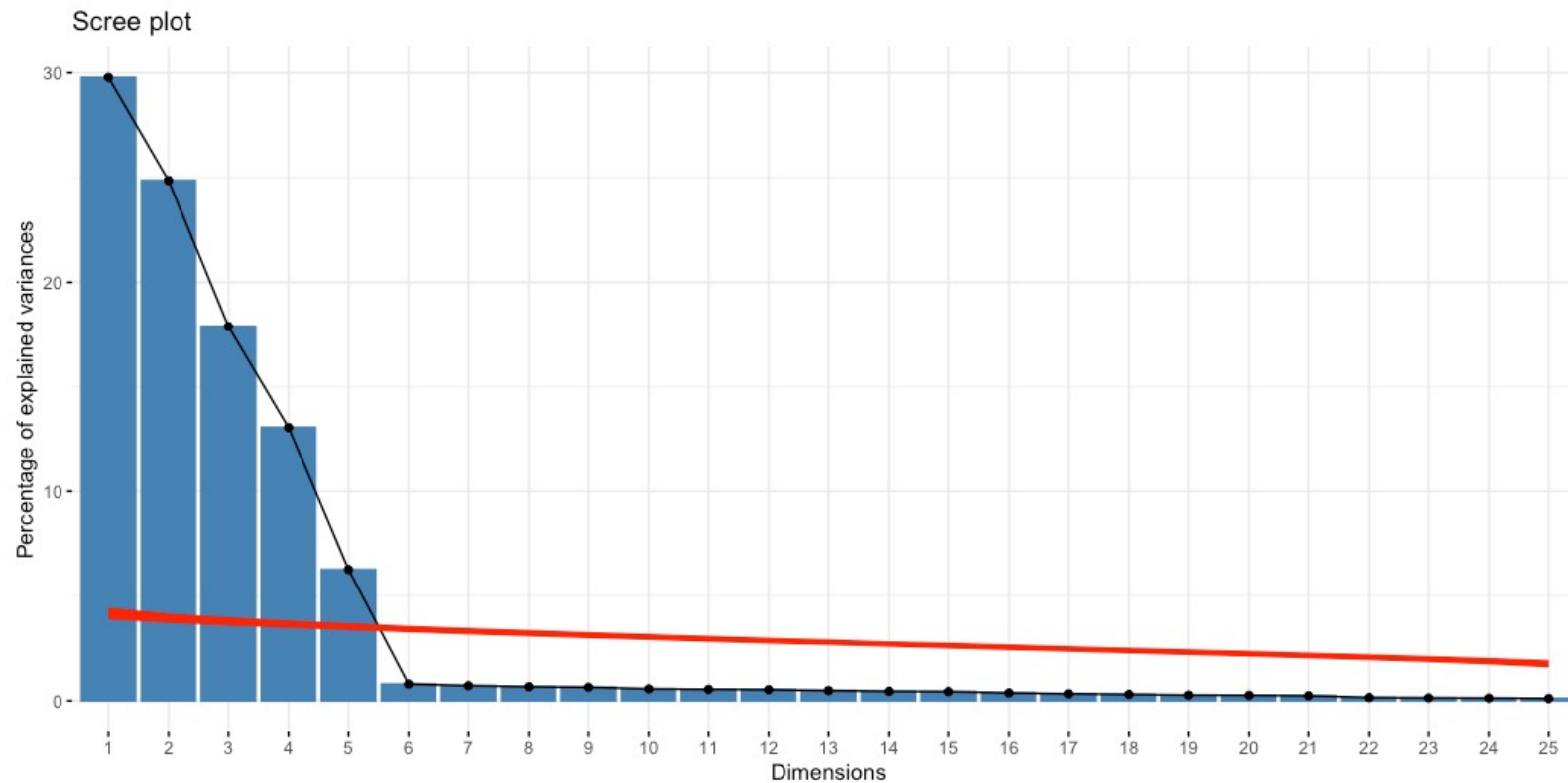
# 4 éboulis caractéristiques



# La règle du coude



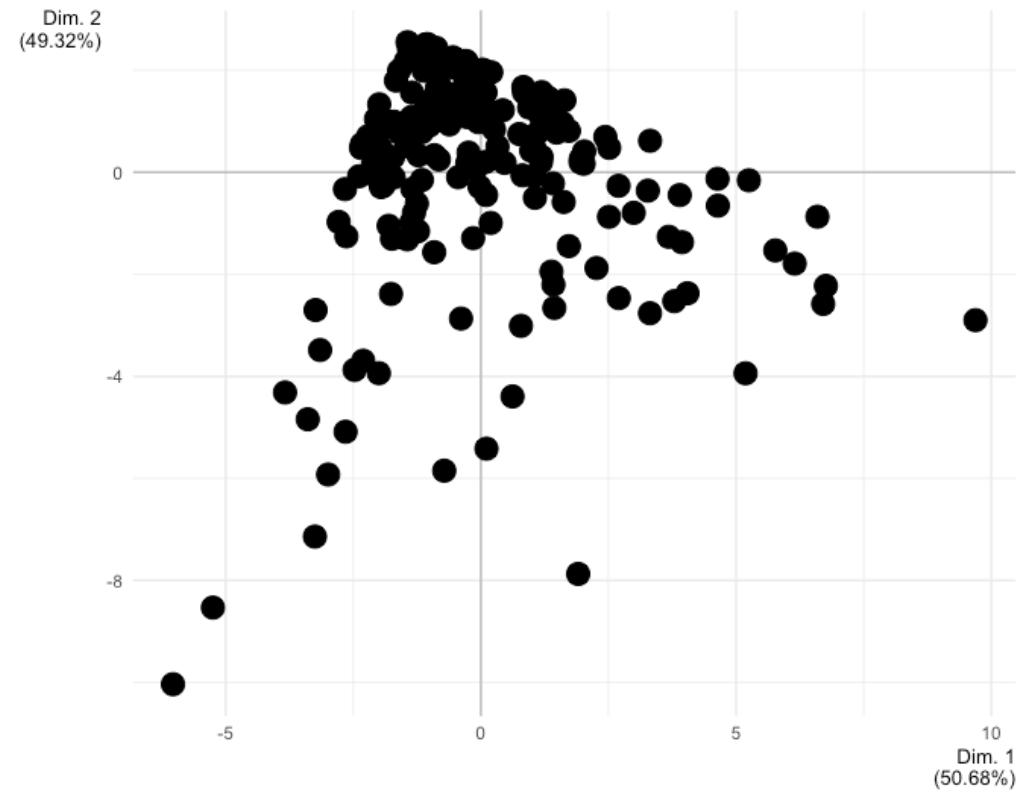
# Les permutations (analyse « parallèle »)



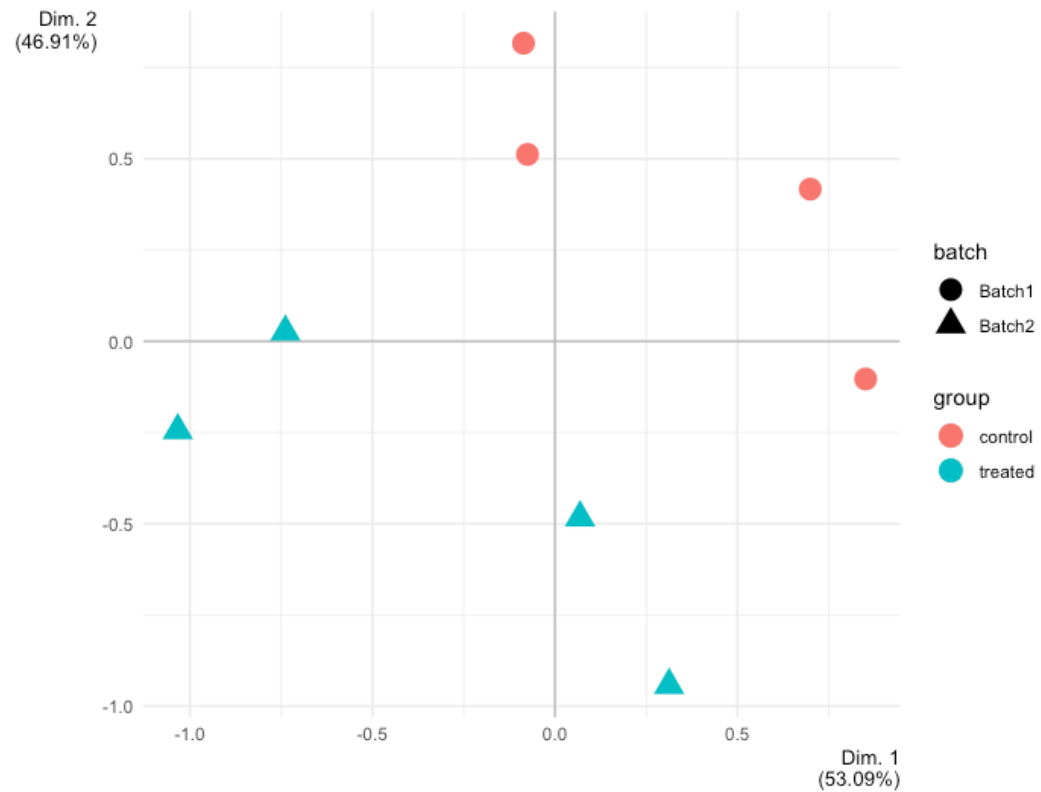
# Une observation isolée



# La carte très dissymétrique

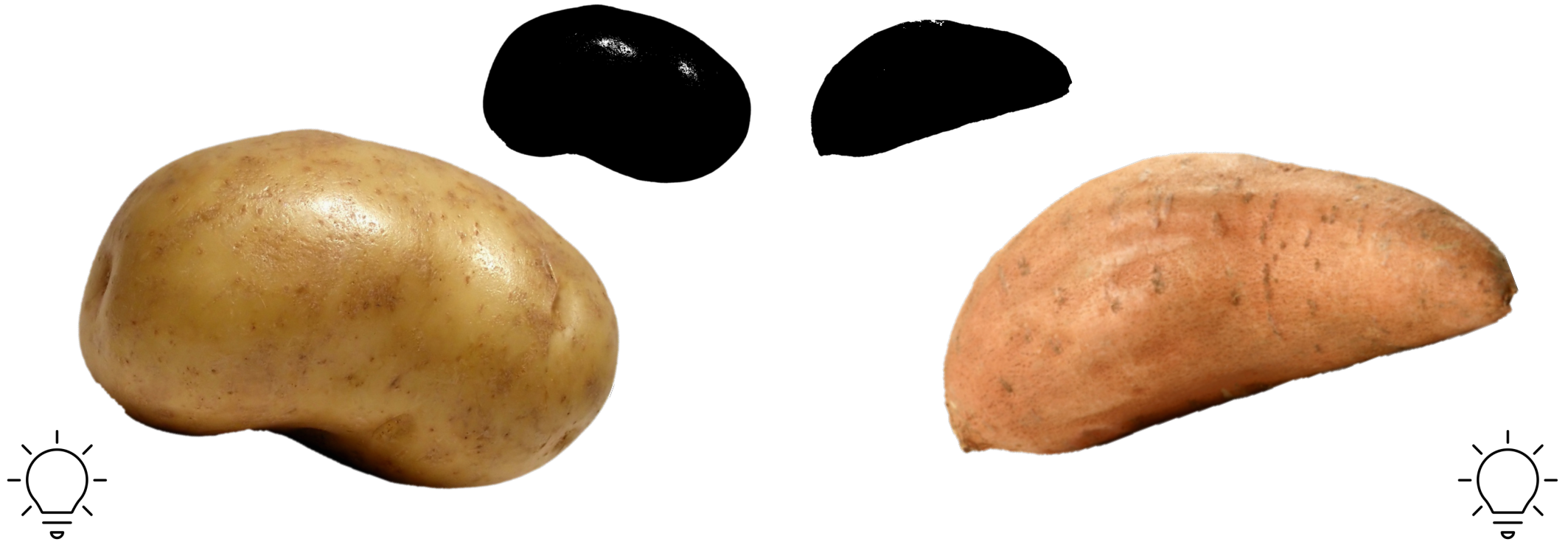


# La confusion

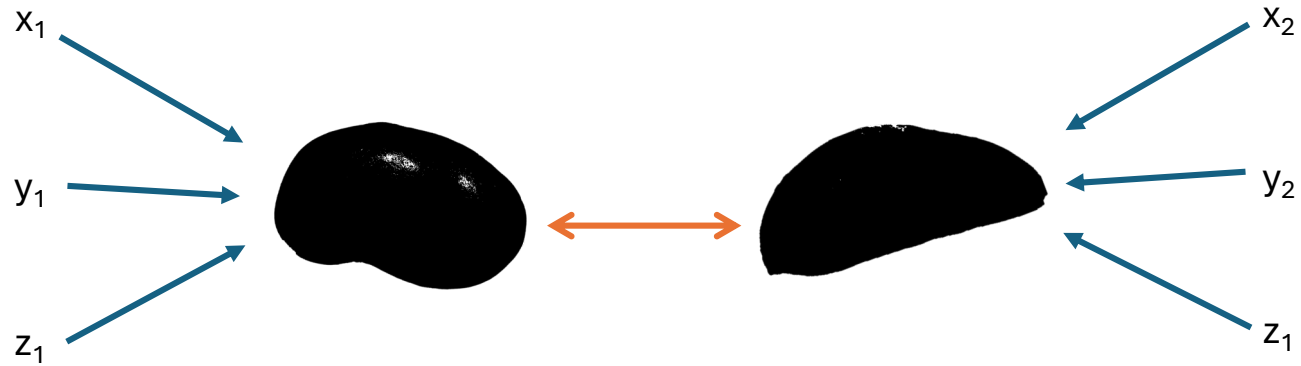


# Analyse Canonique des Corrélations

# Le théâtre des ombres de patates



# Patatomiques



# Résumé

# ACP et ACC sont 2 techniques très similaires

Calculer des coefficients

Calculer des composantes = combinaisons linéaires des variables

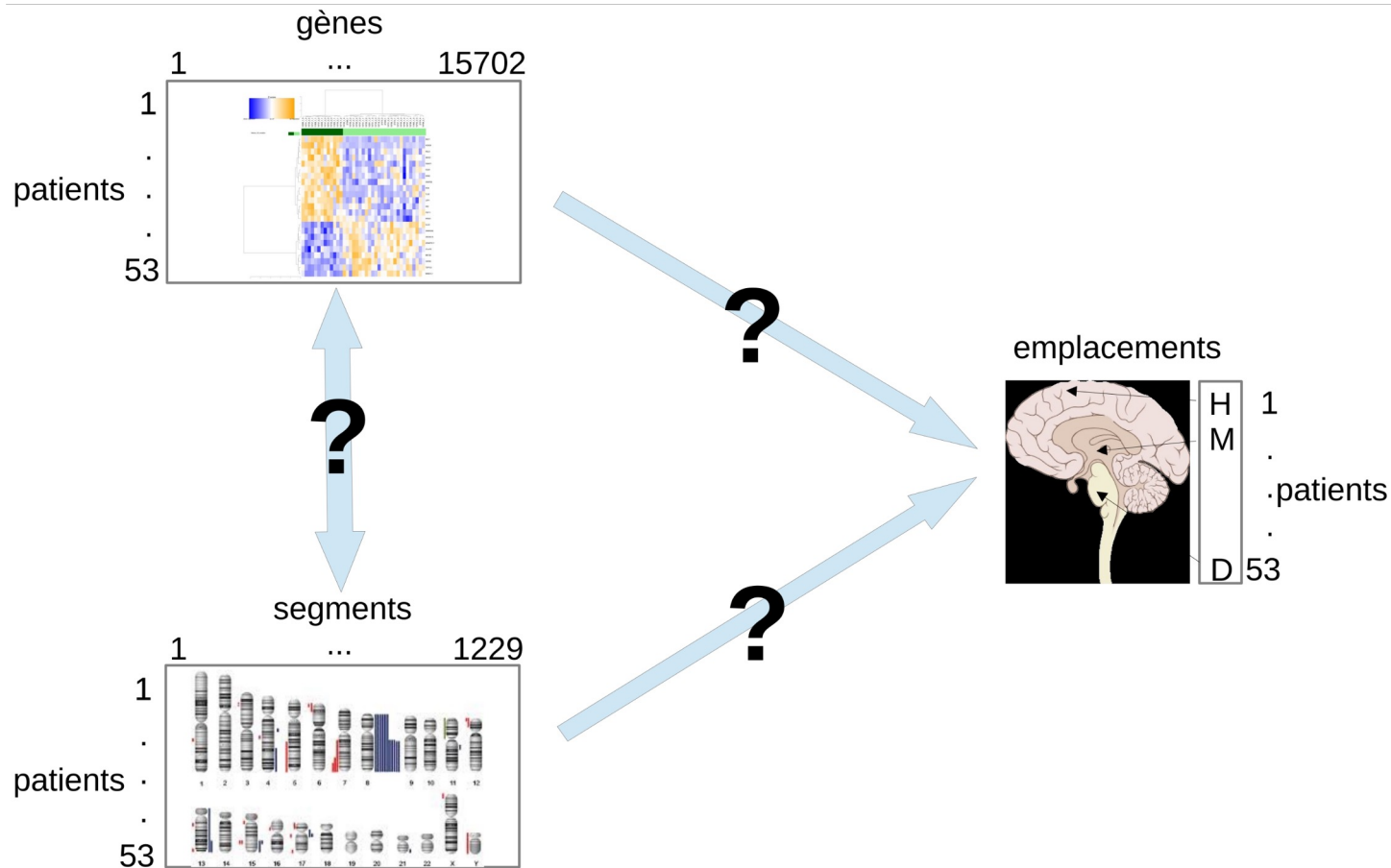
Représenter ces composantes à l'aide de cartes

Visualiser l'importance des variables avec un cercle des corrélations

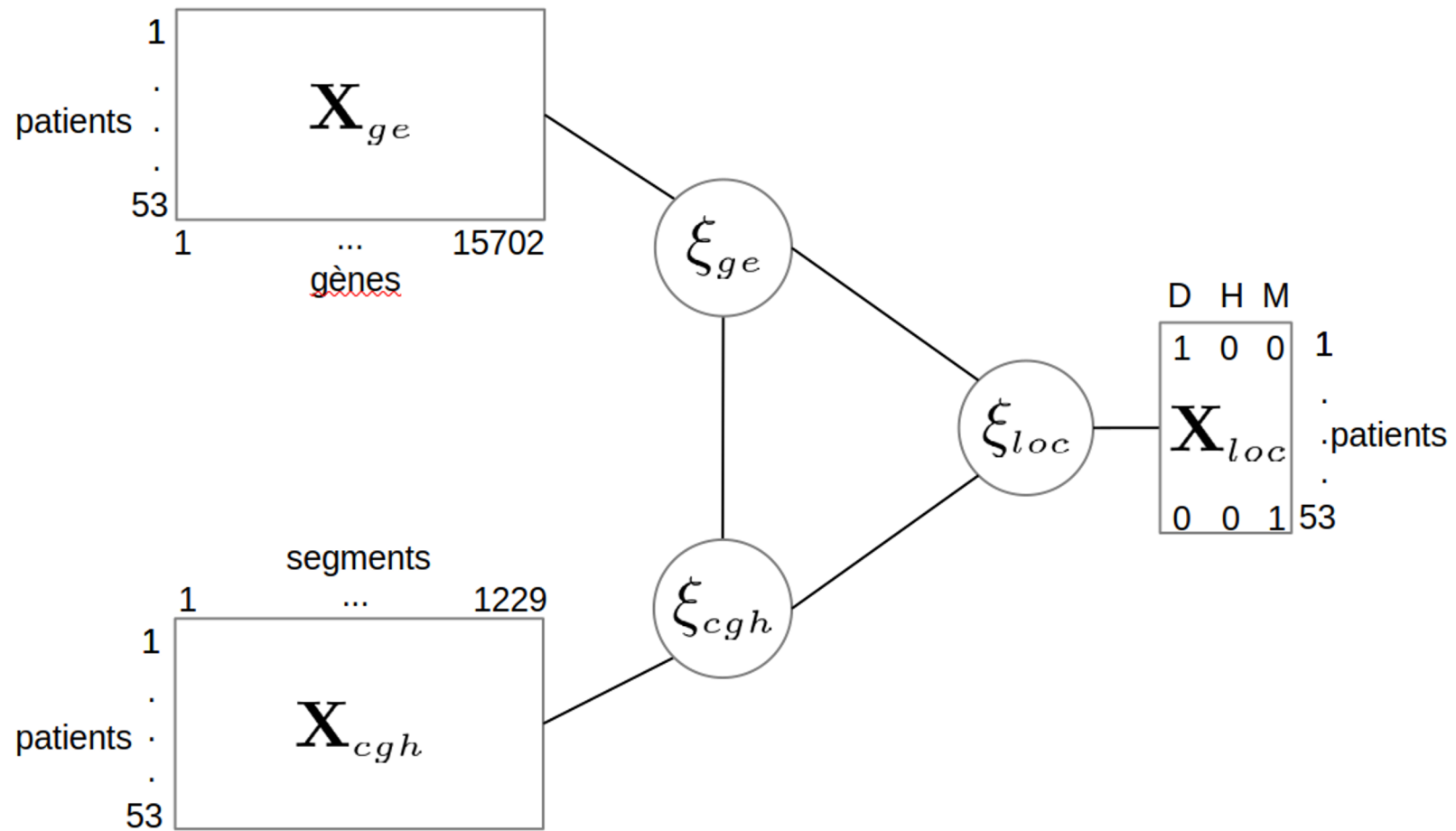
Calculer combien d'information est expliquée par chaque composante

RGCCA

# Le design d'intégration



# Design RGCCA



# Paramètres RGCCA

## Usage

```
rgcca(  
  blocks,  
  connection = NULL,  
  tau = 1,  
  ncomp = 1,  
  scheme = "factorial",  
  scale = TRUE,  
  init = "svd",  
  bias = TRUE,  
  tol = 1e-08,  
  verbose = FALSE,  
  scale_block = "inertia",  
  method = "rgcca",  
  lambda = 0,  
  graph_laplacians = NULL,  
  mu_init = 1,  
  tol_inner = 1e-04,  
  tol_outer = 1e-08,  
  sparsity = 1,  
  response = NULL,  
  superblock = FALSE,  
  NA_method = "na.ignore",  
  quiet = TRUE,  
  n_iter_max = 1000,  
  comp_orth = TRUE,  
  A = NULL,  
  C = NULL,  
  ranks = NULL,  
  woodbury = TRUE  
)
```

## Paramètre blocks

```
class (pHGG)
```

```
[1] "list"
```

```
names (pHGG)
```

```
[1] "cgh" "gex" "loc"
```

## Paramètre connection

```
design = matrix(rep(1, 9), nrow=3, ncol=3)
D = diag(1, 3)
```

```
my_design = design - D
```

```
my_design
```

|       | [, 1] | [, 2] | [, 3] |
|-------|-------|-------|-------|
| [1, ] | 0     | 1     | 1     |
| [2, ] | 1     | 0     | 1     |
| [3, ] | 1     | 1     | 0     |

## Paramètre response

```
...  
response = 3  
...
```

# RGCCA sur gliomaData

```
library(RGCCA)
```

```
my_rgcca_fit = rgcca(blocks = pHGG,  
                      connection = my_design,  
                      response = 3,  
                      ncomp=2)
```

```
print(my_rgcca_fit)
```

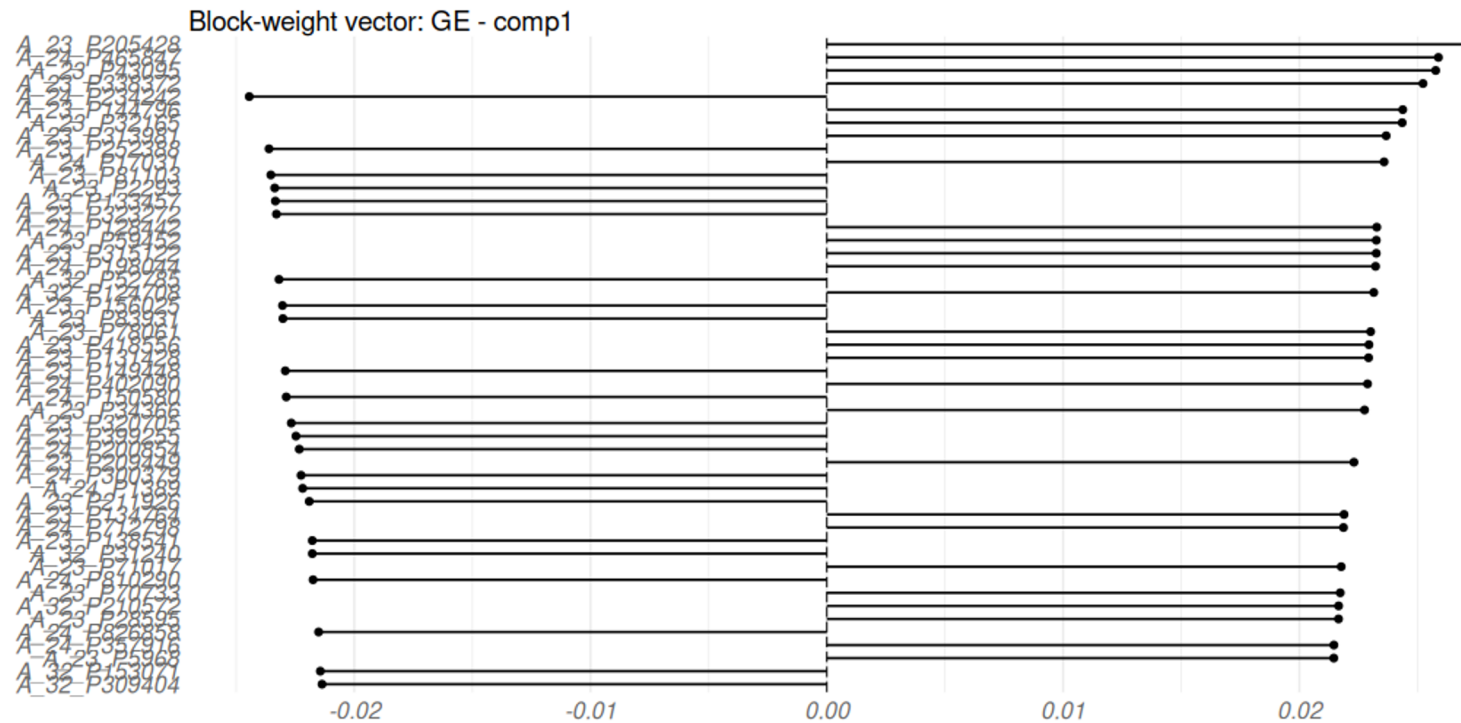
Fitted RGCCA model.

The algorithm converged to a stationnary point:

- After 6 iterations for component 1.
- After 15 iterations for component 2.

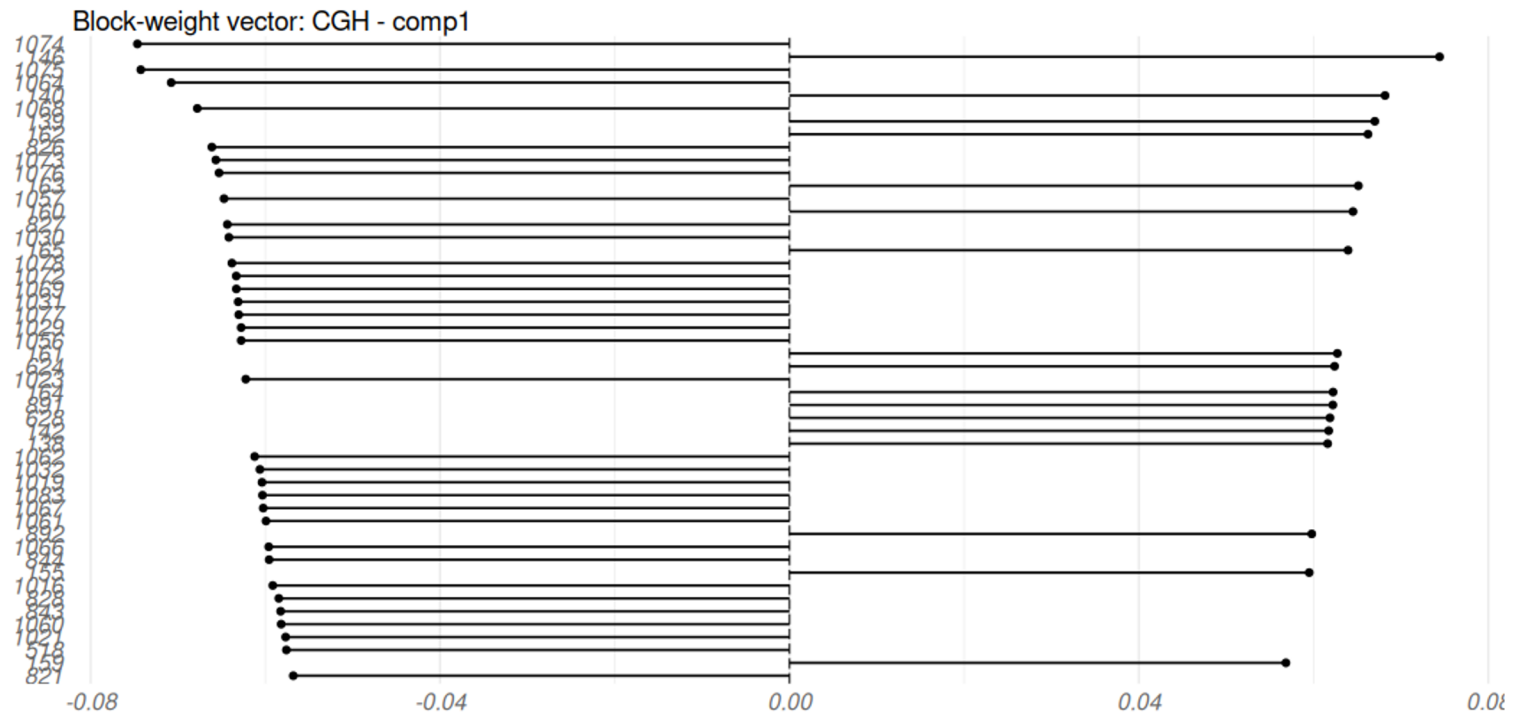
# Graphe des poids block 1

```
plot(my_rgcca_fit, type = "weight", block = 1, display_order = T, n_mark = 50)
```



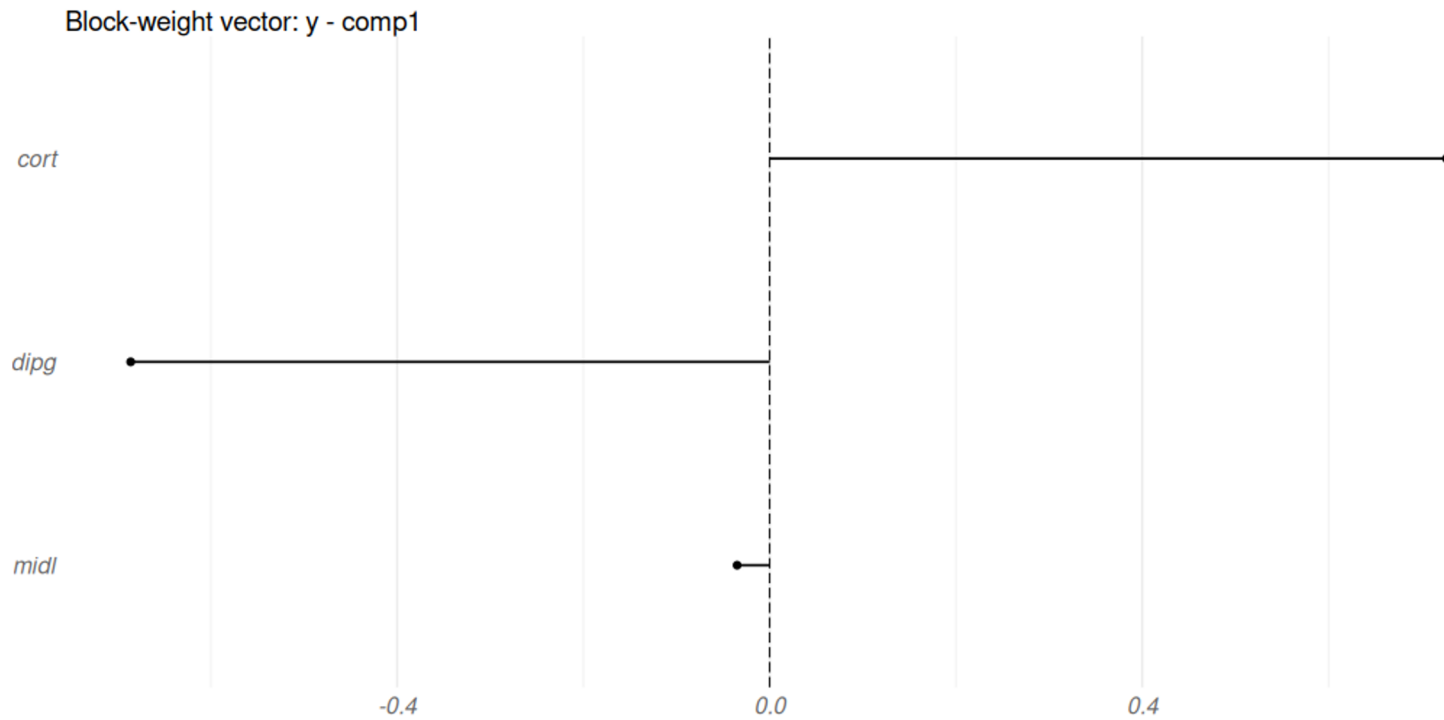
# Graphe des poids block 2

```
plot(my_rgcca_fit, type = "weight", block = 2, display_order = T, n_mark = 50)
```



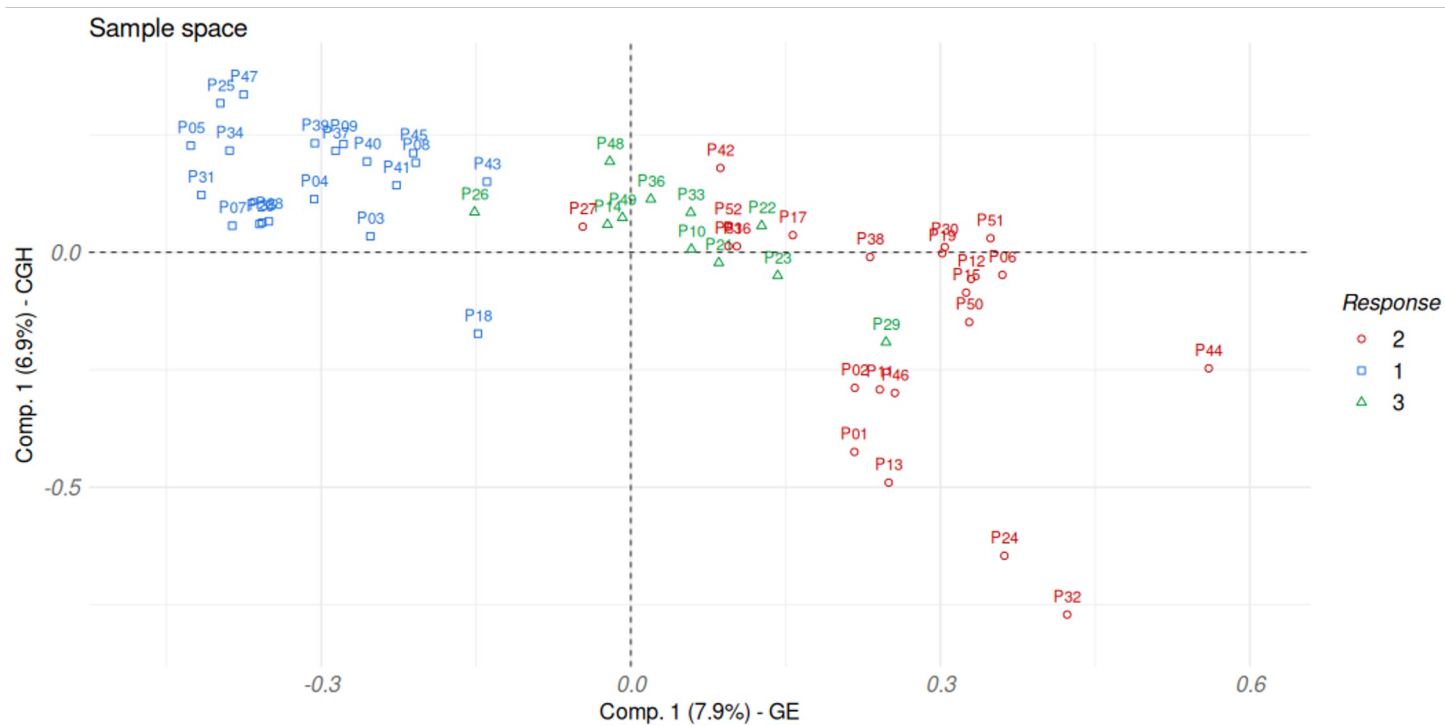
# Graphe des poids block 3

```
plot(my_rgcca_fit, type = "weight", block = 3, display_order = T)
```



# Graphes des individus

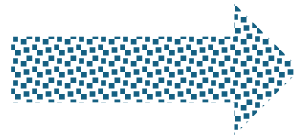
```
plot(my_rgcca_fit,  
     type = "sample", block = 1:2,  
     comp = rep(1, 2), resp = as.factor(pHGG$loc)  
)
```



**SGCCA**

# Parcimonie

| Gene   | Coefficient |
|--------|-------------|
| IL1A   | 0,5         |
| IFNG   | -3          |
| CEA    | 2           |
| INSERM | 12          |



| Gene   | Coefficient  |
|--------|--------------|
| IL1A   | X            |
| IFNG   | négatif      |
| CEA    | X            |
| INSERM | très positif |

# Activité

Réfléchissez à une procédure qui permette de rendre un vecteur de coefficient « parcimonieux ». Si vous vous sentez suffisamment à l'aide, faites-en une fonction R !

*Exemple* : je sélectionne au hasard des coefficients que je mets à zéro, les autres coefficients sont inchangés

```
random_sparsity <- function(coefs, nb) {  
  out <- coefs  
  out[sample(length(out), nb)] <- 0  
  return(out)  
}
```

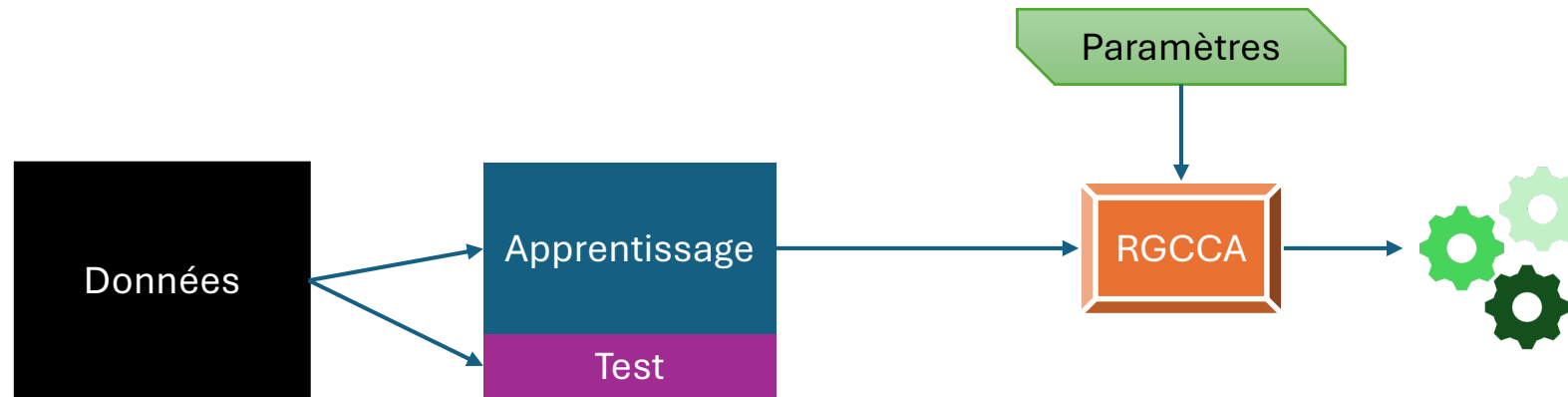
# Démonstration

```
res.sgcca <- rgcca(  
  blocks = pHGG,  
  sparsity = c(0.1, 0.1, 1),  
  response = 3)
```

On applique SGCCA (== RGCCA avec parcimonie) sur les données pHGG :

- avec un coefficient de parcimonie petit => beaucoup de zéros
- avec un coefficient de parcimonie moyen => entre les deux
- avec un coefficient de parcimonie grand => pas beaucoup de zéros

# Validation croisée



| Paramètre | Performance |
|-----------|-------------|
| Valeur 1  | Accuracy 1  |
| Valeur 2  | Accuracy 2  |
| Valeur 3  | Accuracy 3  |

# Conclusion

- RGCCA est un package
  - versatile : MCIA, MFA, CCA, PCA etc.
  - complet : CV, bootstrap, parcimonie
- Mais ça ne fait pas tout !!!!
- Le plus important, c'est de réfléchir à la question biologique

