13/12/2022

Projet DataViz

Sur des données génomiques issues du “The Cancer Génome Atlas”

Lama Zidan
MASTER 1- UNIVERSITE LYON2

Table des matières

Présentation du contexte et de la base de données et des tâches retenues :

Nous disposons d'un jeu de données qui représente les données génomiques de patients atteints de cancer avec données de survie associées Ces données sont issues du "The Cancer Génome Atlas" (TCGA). Nous avons 1456 observations(les patients du cancer) et 719 variables.

Les premières quatre variables sont

- Row.names qui indique le code du patient concerné,
- CLINIC.OS.MONTHS qui est la durée de vie en mois donc quantitative continue
- CLINIC.OS_STATUS il nous informe de l'état actuel du patient, s'il est décédé ou vivant(0 pour LIVING et 1 pour DECEASED) donc qualitative binaire.
- CLINIC.study c'est le type de cancer correspondant donc qualitative de trois modalités prad pour le cancer de Prostate, paad pour le cancer de Pancréas et brca pour le cancer de Sein.
- Et 715 autres variables quantitatives continues normalisées qui sont l'expression des 715 gènes pour chaque patient.

Comme les expressions des gènes (les variables) ne sont pas très parlants et nous ne savons pas comment ils fonctionnent ou de quelle manière ils varient et sous quelles conditions, etc.. l'objectif de ce projet va être donc d'essayer d'étudier du près ces gènes pour extraire des informations qui pourraient être pertinents. En d'autres termes on va essayer de découvrir les gènes les plus importants qui nous permet de déterminer le type de cancer atteint et d'étudier la corrélation entre eux et de représenter graphiquement les corrélation correspondants.

Donc dans notre étude on s'intéresse à la variable CLINIC.study qui contient trois types de cancers comme variable cible et les 715 autres variables (les gènes) pour essayer de comprendre quel relation existe entre expression des gènes et le type de cancer atteint. C'est-à-dire essayer d'extraire des informations nous indiquant quel changement d'expression génétique est le responsable le plus d'entrainer un type de cancer donné.

Les étapes de notre études :

1. Analyse générale et prétraitement du jeu de données, nous allons découvrir et analyser les données de façon générale, ainsi que leurs distribution et mettre en place plusieurs représentations graphiques (matrice de corrélation, histogramme, distribution, etc..).
2. Développement d'une étude de la société canadienne du cancer, nous allons se baser sur cet étude pour ensuite aller plus loin dans notre analyse sur les gènes.
3. ACP, nous allons appliquer l'ACP afin de voir si les données en des dimensions réduites formeront des clusters ou non, et représenter plusieurs graphiques intéressantes.
4. Conclusion et rediscuter des résultats observés.

1.Analyse générale et prétraitement du jeu de données:

Affichons les colonnes de notre base :

```
: #afficher le nom  des colonnes:
  df.columns
```

```
: Index(['Row.names', 'CLINIC.OS_MONTHS', 'CLINIC.OS_STATUS', 'CLINIC.study',
         'EXP.A1CF', 'EXP.ABI1', 'EXP.ABL1', 'EXP.ABL2', 'EXP.ACKR3',
         'EXP.ACSL3',
         ...
         'EXP.ZFHX3', 'EXP.ZMYM2', 'EXP.ZMYM3', 'EXP.ZNF331', 'EXP.ZNF384',
         'EXP.ZNF429', 'EXP.ZNF479', 'EXP.ZNF521', 'EXP.ZNRF3', 'EXP.ZRSR2'],
        dtype='object', length=719)
```

Notre jeu de données se rassemble à ceci :

```
Entrée [101]: df.head()
              #les données sont deja normalisées
```

Out[101]:

	Row.names	CLINIC.OS_MONTHS	CLINIC.OS_STATUS	CLINIC.study	EXP.A1CF	EXP.ABI1	EXP.ABL1	EXP.ABL2	EXP.ACKR3	EXP.ACSL3	...	EXP.ZFH
1	TCGA.2A.A8VL.01	NaN	NaN	prad	0.000000	0.541512	0.604870	0.482775	0.347875	0.712888	...	0.5939
2	TCGA.2A.A8VO.01	55.88	0:LIVING	prad	0.000000	0.547076	0.586030	0.483318	0.405532	0.657795	...	0.5780
3	TCGA.2A.A8VT.01	45.11	0:LIVING	prad	0.024106	0.528198	0.608390	0.533721	0.398137	0.707258	...	0.5757
4	TCGA.2A.A8VV.01	22.04	0:LIVING	prad	0.000000	0.560709	0.618023	0.488945	0.335193	0.712310	...	0.5971
5	TCGA.2A.A8VX.01	45.27	0:LIVING	prad	0.000000	0.559985	0.589078	0.491609	0.374201	0.647905	...	0.5987

5 rows × 719 columns

Les dimensions de notre base : 719 variables et 1456 observations, on peut dire que le nombre d'observation est deux fois plus le nombre des variables.

```
ntrée [66]: df.shape
```

```
Out[66]: (1456, 719)
```

Nous allons tout d'abord supprimer les colonnes qui ne nous intéressent pas :

CLINIC.OS_STATUS , CLINIC.OS_MONTHS, Row.names pour se focaliser sur notre variable cible 'CLINIC.study' et sa relation avec les 715 gènes.

```
Entrée [104]: #on va s'interesser dans ce projet à la variable cible CLINIC.study
              #donc on va supprimer les variables CLINIC.OS_STATUS et CLINIC.OS_MONTHS
              df=df.drop(['CLINIC.OS_STATUS','CLINIC.OS_MONTHS','Row.names'], axis=1)

              df.head()
```

Out[104]:

	CLINIC.study	EXP.A1CF	EXP.ABI1	EXP.ABL1	EXP.ABL2	EXP.ACKR3	EXP.ACSL3	EXP.ACSL6	EXP.ACVR1	EXP.ACVR2A	...	EXP.ZFHX3	EXP.ZMYM2	EXP.
1	prad	0.000000	0.541512	0.604870	0.482775	0.347875	0.712888	0.125749	0.500824	0.457611	...	0.593948	0.582641	0.
2	prad	0.000000	0.547076	0.586030	0.483318	0.405532	0.657795	0.208305	0.462957	0.472713	...	0.578088	0.570839	0.
3	prad	0.024106	0.528198	0.608390	0.533721	0.398137	0.707258	0.266585	0.533194	0.463842	...	0.575769	0.623860	0.
4	prad	0.000000	0.560709	0.618023	0.488945	0.335193	0.712310	0.134106	0.494678	0.470423	...	0.597133	0.593029	0.
5	prad	0.000000	0.559985	0.589078	0.491609	0.374201	0.647905	0.053670	0.460583	0.496246	...	0.598723	0.551116	0.

5 rows × 716 columns

Nous avons maintenant nos features, les 715 expression des gènes et la variable cible CLINIC.study qui représente le type de cancer.

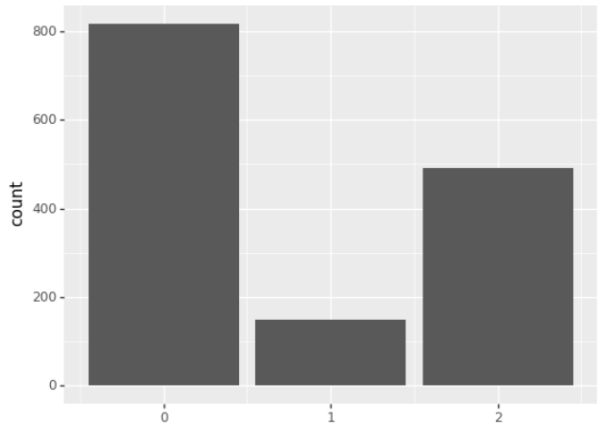
Nous remarquons que notre base de données ne contient aucun valeur manquante :

```
Entrée [69]: #missing data
             #on a pas de valeurs manquantes
             # Nombre total de valeurs manquantes
             print(df.isnull().sum().sum())

             0
```

Nous intéressons maintenant à connaitre le nombre d'observation pour chaque type de cancer :On précise bien ici que prad est le cancer de la prostate, paad est le cancer de la pancréas

Et brca est le cancer du sein.

```
Entrée [170]: #nombre d'observation pour chaque type de cancer:
              df['CLINIC.study'].value_counts()
              #prad=prostate
              #paad=pancreas
              #brca=Sein

Out[170]: 0.0    816
          2.0    491
          1.0    149
          Name: CLINIC.study, dtype: int64

Entrée [171]: #histogramme d'observation pour chaque type de cancer:


              (ggplot(df, aes(x='CLINIC.study')) + \
                  geom_bar(stat = 'count')
              )
```



On remarque bien que le cancer de Sein est le plus présente dans notre jeu de données ce qui est presque deux fois plus le nombre d'observation du cancer de Prostate et presque 5 fois plus du cancer de Pancréas.

Nous allons nous s'intéresser maintenant à la distribution de nos données . Pour cela on enregistre les features dans une variable X. Nous remarquons bien que les données sont déjà normalisées, on affiche leurs description :

Entrée [108]:
```
X=df.drop(['CLINIC.study'], axis=1)
X
```

Out[108]:

	EXP.A1CF	EXP.ABI1	EXP.ABL1	EXP.ABL2	EXP.ACKR3	EXP.ACSL3	EXP.ACSL6	EXP.ACVR1	EXP.ACVR2A	EXP.AFDN	...	EXP.ZFHX3	EXP.ZMYM2	EXP.
1	0.000000	0.541512	0.604870	0.482775	0.347875	0.712888	0.125749	0.500824	0.457611	0.653970	...	0.593948	0.582641	0
2	0.000000	0.547076	0.586030	0.483318	0.405532	0.657795	0.208305	0.462957	0.472713	0.610044	...	0.578088	0.570839	0
3	0.024106	0.528198	0.608390	0.533721	0.398137	0.707258	0.266585	0.533194	0.463842	0.635984	...	0.575769	0.623860	0
4	0.000000	0.560709	0.618023	0.488945	0.335193	0.712310	0.134106	0.494678	0.470423	0.626111	...	0.597133	0.593029	0
5	0.000000	0.559985	0.589078	0.491609	0.374201	0.647905	0.053670	0.460583	0.496246	0.622077	...	0.598723	0.551116	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1452	0.000000	0.559250	0.623345	0.516899	0.400225	0.588526	0.240718	0.550710	0.464334	0.627647	...	0.516619	0.583308	0
1453	0.000000	0.558034	0.612417	0.500689	0.461815	0.590283	0.181130	0.539011	0.455220	0.615302	...	0.561513	0.579024	0
1454	0.000000	0.551856	0.593140	0.525267	0.401069	0.608822	0.161938	0.485093	0.480452	0.608120	...	0.512136	0.596186	0
1455	0.000000	0.532657	0.618559	0.485027	0.401104	0.653819	0.102120	0.429514	0.443638	0.637753	...	0.534660	0.530551	0
1456	0.346859	0.538288	0.616085	0.473884	0.460049	0.650160	0.070649	0.471891	0.494001	0.530133	...	0.564359	0.518039	0

1456 rows × 715 columns

Entrée [109]:
```
X.describe()
```

Out[109]:

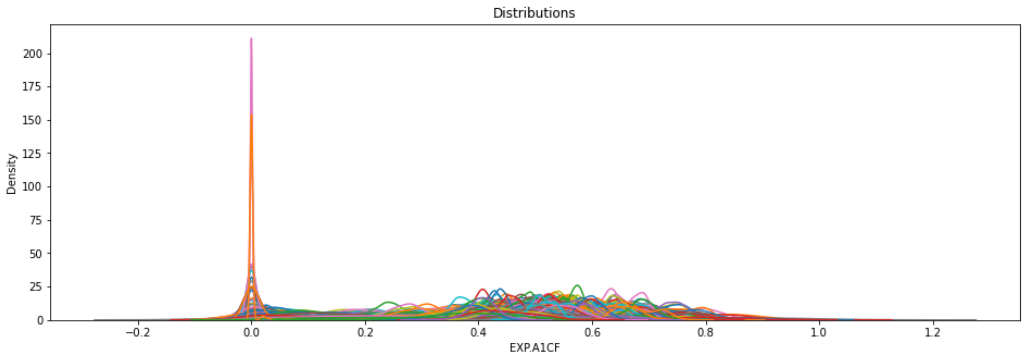
	EXP.A1CF	EXP.ABI1	EXP.ABL1	EXP.ABL2	EXP.ACKR3	EXP.ACSL3	EXP.ACSL6	EXP.ACVR1	EXP.ACVR2A	EXP.AFDN	...	EXP.ZFHX3	E
count	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	...	1456.000000	1
mean	0.037833	0.532702	0.568331	0.503167	0.425122	0.611159	0.167227	0.491077	0.432410	0.578563	...	0.512578	
std	0.085857	0.020385	0.035233	0.025438	0.061353	0.060148	0.074171	0.027944	0.040318	0.051388	...	0.059438	
min	0.000000	0.422011	0.463623	0.409862	0.202742	0.466746	0.000000	0.372165	0.245780	0.000000	...	0.298877	
25%	0.000000	0.520116	0.543055	0.486747	0.386290	0.566993	0.118381	0.473830	0.404353	0.545182	...	0.472418	
50%	0.000000	0.532604	0.562313	0.503252	0.425386	0.600013	0.159049	0.493673	0.435454	0.576210	...	0.502954	
75%	0.023900	0.545170	0.598918	0.520271	0.465656	0.653619	0.204438	0.510115	0.463380	0.614032	...	0.557724	
max	0.421644	0.645653	0.671399	0.598201	0.662149	0.792991	0.531884	0.581760	0.545464	0.721779	...	0.660658	

8 rows × 715 columns

Les expressions des gènes varient entre 0 et 1.

Nous affichons la distribution de tous nos features:

Entrée [28]:
```
fig, a = plt.subplots(ncols=1, figsize=(16, 5))
a.set_title("Distributions")
for col in X.columns:
    sns.kdeplot(X[col], ax=a)
plt.show()
```

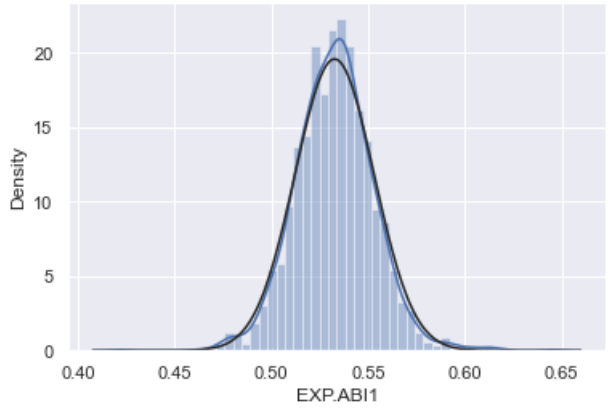


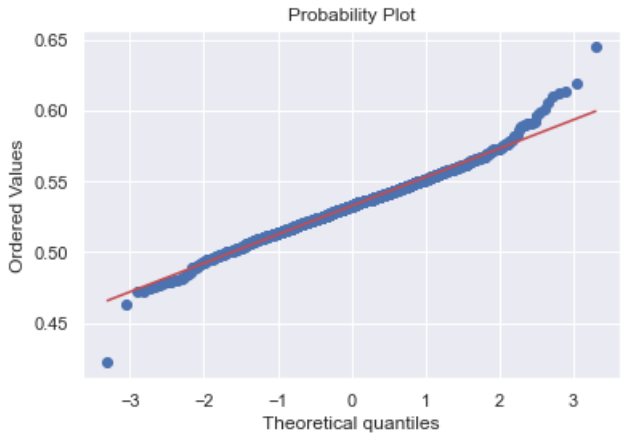
On peut dire que, plusieurs gènes ont un pic dans le point 0 et qui ont un aplatissement très élevée (jusqu'au 200) mais la majorité des gènes ont leurs pic après le 0 et leurs aplatissement est faible(jusqu'au 25).

On décide de prendre un gène au hasard et voir sa distribution :

Entrée [76]:
```
#histogram and normal probability plot
sns.distplot(df['EXP.ABI1'], fit=norm);
fig = plt.figure()
res = stats.probplot(df['EXP.ABI1'], plot=plt)
print("Skewness: %f" % df['EXP.ABI1'].skew())
print("Kurtosis: %f" % df['EXP.ABI1'].kurt())
```

```
Skewness: 0.184059
Kurtosis: 1.907057
```





On remarque bien que sa distribution est tout à fait normale !

Avec un Skewness 0.18, nous pouvons dire que la distribution est normale puisque entre -0.5 et 0.5 la distribution est considéré normale.

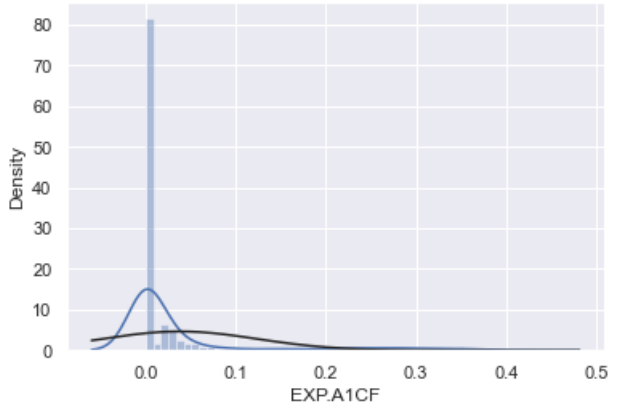
Et un Kurtosis (aplatissement) de 1.9 presque normale.

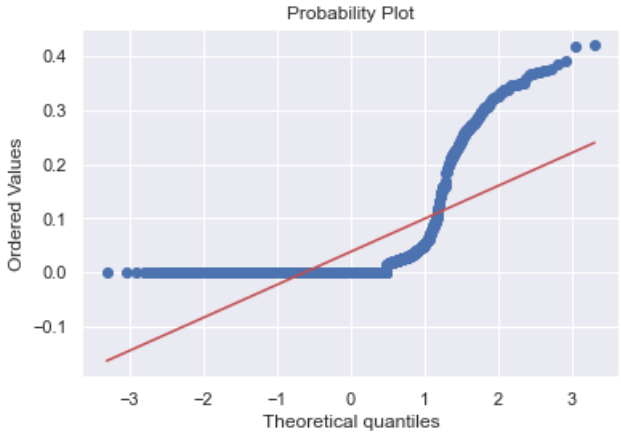
On prend un autre gène :

Entrée [77]:

```
#histogram and normal probability plot
sns.distplot(df['EXP.A1CF'], fit=norm);
fig = plt.figure()
res = stats.probplot(df['EXP.A1CF'], plot=plt)
print("Skewness: %f" % df['EXP.A1CF'].skew())
print("Kurtosis: %f" % df['EXP.A1CF'].kurt())
```

```
Skewness: 2.541094
Kurtosis: 5.370337
```





cette fois c'est le contraire puisque la distribution n'est pas du tout normale.

Regardons la matrice de corrélations de nos gènes et le Heatmap correspondant :

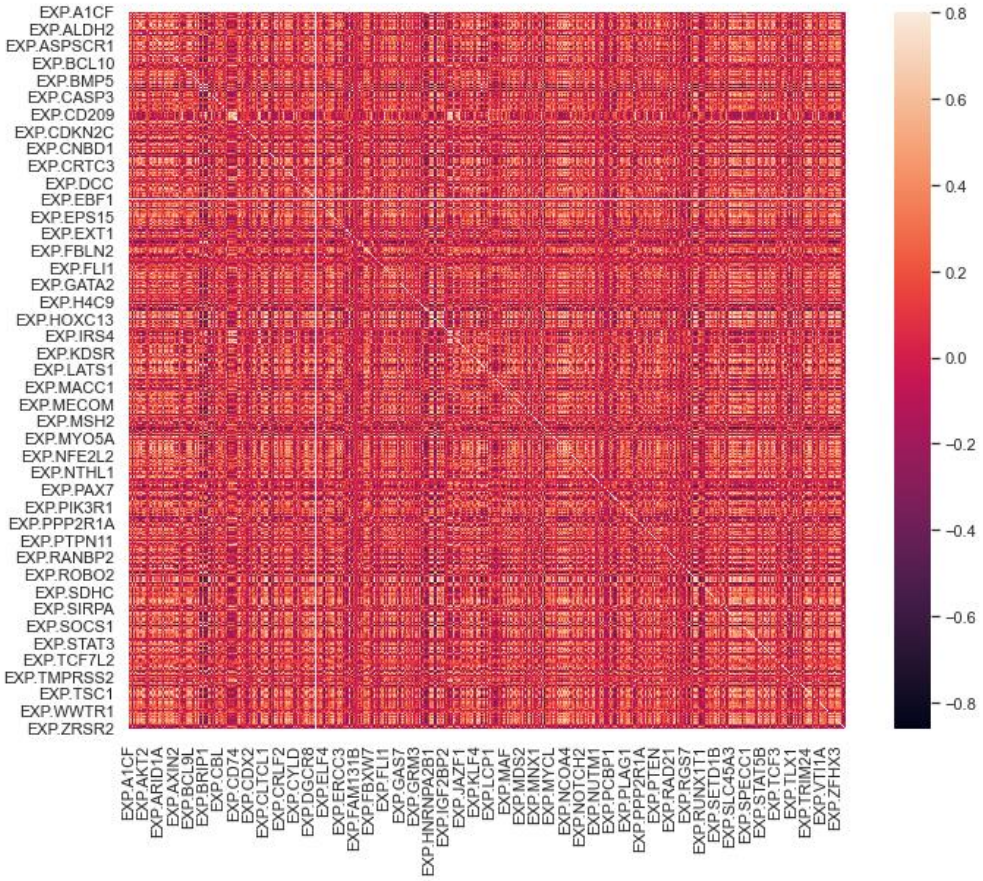
Entrée [83]:
```
#afficher la matrice de correlation entre les genes:
pw_corr=pd.DataFrame(X).corr().round(3)
print(pw_corr)
```

```
            EXP.A1CF  EXP.ABI1  EXP.ABL1  EXP.ABL2  EXP.ACKR3  EXP.ACSL3  \
EXP.A1CF       1.000     0.062    -0.011    -0.094      0.125     -0.253
EXP.ABI1       0.062     1.000     0.230     0.168     -0.144      0.281
EXP.ABL1      -0.011     0.230     1.000     0.034     -0.217      0.587
EXP.ABL2      -0.094     0.168     0.034     1.000      0.229     -0.064
EXP.ACKR3      0.125    -0.144    -0.217     0.229      1.000     -0.388
...              ...       ...       ...       ...        ...        ...
EXP.ZNF429    -0.333     0.076     0.290     0.095     -0.207      0.456
EXP.ZNF479    -0.021    -0.032    -0.056    -0.006      0.024     -0.001
EXP.ZNF521    -0.008    -0.036    -0.346     0.366      0.460     -0.478
EXP.ZNRF3     -0.073     0.214     0.578    -0.023     -0.274      0.598
EXP.ZRSR2     -0.010    -0.239    -0.044    -0.228     -0.087     -0.116

            EXP.ACSL6  EXP.ACVR1  EXP.ACVR2A  EXP.AFDN  ...  EXP.ZFHX3  \
EXP.A1CF       -0.015      0.206       0.168    -0.028  ...     -0.048
EXP.ABI1        0.140      0.256       0.324     0.417  ...      0.318
EXP.ABL1       -0.076      0.250       0.590     0.528  ...      0.711
EXP.ABL2        0.241      0.187      -0.088     0.072  ...     -0.000
EXP.ACKR3       0.071      0.257      -0.160    -0.286  ...     -0.243
...               ...        ...         ...       ...  ...        ...
EXP.ZNF429     -0.073     -0.011       0.187     0.321  ...      0.445
EXP.ZNF479     -0.002     -0.012      -0.052    -0.053  ...     -0.054
EXP.ZNF521      0.256      0.239      -0.221    -0.348  ...     -0.419
EXP.ZNRF3      -0.002      0.034       0.401     0.551  ...      0.681
EXP.ZRSR2      -0.098     -0.218       0.013    -0.191  ...     -0.105

            EXP.ZMYM2  EXP.ZMYM3  EXP.ZNF331  EXP.ZNF384  EXP.ZNF429  \
EXP.A1CF       -0.086     -0.177      -0.043      -0.014      -0.333
EXP.ABI1        0.377      0.234       0.215       0.206       0.076
EXP.ABL1        0.558      0.639       0.639       0.452       0.290
EXP.ABL2        0.160     -0.005      -0.041      -0.060       0.095
EXP.ACKR3      -0.262     -0.347      -0.257      -0.301      -0.207
...               ...        ...         ...         ...         ...
EXP.ZNF429      0.507      0.428       0.330       0.193       1.000
EXP.ZNF479     -0.043     -0.058      -0.095      -0.040       0.064
EXP.ZNF521     -0.325     -0.363      -0.402      -0.230      -0.216
EXP.ZNRF3       0.492      0.543       0.588       0.388       0.347
EXP.ZRSR2      -0.195     -0.018       0.071       0.014      -0.015
```

Nous pouvons remarquer la corrélation qui pourrait être négative et positive entre les gènes et qui est assez varié en terme de chiffres !

Affichons le Heatmap correspondant:

```
Entrée [26]: corrmat = X.corr()
             f, ax = plt.subplots(figsize=(12, 9))
             sns.heatmap(corrmat, vmax=.8, square=True);
```



Nous pouvons remarquer l'existence des corrélation négatives et positives en suivant les couleurs de notre Heatmap, nous remarquons des champs claires et sombres et bien des corrélations simplement neutres.

Comme cette représentation ne nous permet pas de dire grandes choses à cause de l'illisibilité, nous allons s'aider d'une article qui pourrait nous dire plus sur ces gènes et leurs relations avec le type de cancer atteints et nous mettre sur le bon chemin.

2. Développement d'une étude de la société canadienne du cancer :

Lien vers l'article :(Mutations des gènes BRCA :selon la société canadienne du cancer : Changements génétiques et risque de cancer | Société canadienne du cancer).

On cite une partie de l'étude qui pourrait nous intéresser dans ce projet :

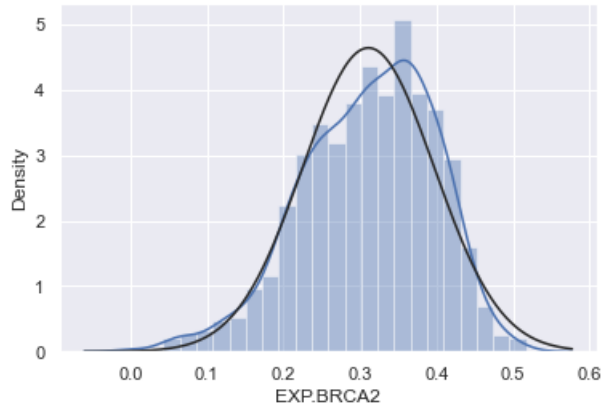
« Les gènes BRCA sont des gènes suppresseurs de tumeurs qui aident habituellement à prévenir l'apparition du cancer. Ils contrôlent la croissance et la division des cellules et aident à la réparation des dommages causés à l'ADN. Toutefois, des gènes BRCA qui ont subi une mutation peuvent accroître le risque d'apparition de certains types de cancer. Il existe 2 mutations des gènes BRCA qui sont connues pour provoquer le cancer – la mutation du gène BRCA1 et la mutation du gène BRCA2. Ces mutations génétiques(BRCA1 et BRCA2) font augmenter le risque d'une femme d'être un jour atteinte d'un cancer du sein et d'un cancer de l'ovaire. Les mutations du gène BRCA2 sont aussi liées à une hausse du risque de cancer du sein et du cancer de la prostate chez l'homme. Les mutations du gène BRCA2 rendent aussi l'homme et la femme un peu plus susceptibles d'être un jour atteints d'un cancer du pancréas. »

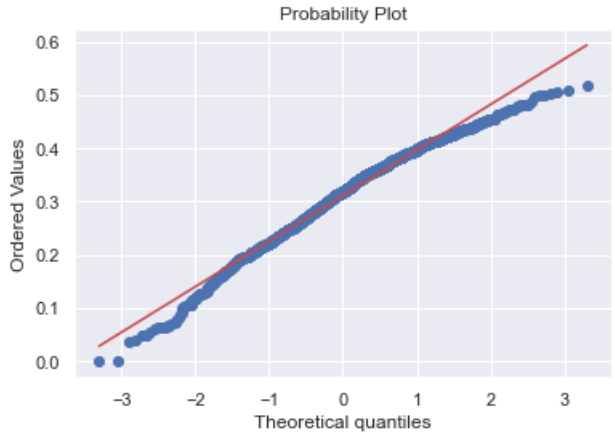
On en conclut que les mutations du gène BRCA2 vont entrainer des changement dans l'expression de plusieurs gènes qui va en conséquence augmenter le risque d'atteindre le cancer du Prostate ou de Pancréas ou de Sein. Comme notre jeu de données contient bien ces trois types de cancer, nous allons essayer d'extraire des informations qui pourrait être intéressantes à partir de cette étude.

On se concentre alors sur le gène BRCA2, présentons d'abord sa distribution :

Entrée [86]:
```
#histogram and normal probability plot
sns.distplot(df['EXP.BRCA2'], fit=norm);
fig = plt.figure()
res = stats.probplot(df['EXP.BRCA2'], plot=plt)
print("Skewness: %f" % df['EXP.BRCA2'].skew())
print("Kurtosis: %f" % df['EXP.BRCA2'].kurt())
```

```
Skewness: -0.473082
Kurtosis: 0.036200
```





on voit que sa distribution est presque normale.

Nous souhaitons vérifier l'étude cité plus haut en répondant à la question suivante :

Est-ce qu'il y a une corrélation entre ce gène et le type de cancer ? c'est à dire est-ce que son expression varie remarquablement pour chaque type de cancer ?

Pour répondre à cette question, nous allons utiliser le test statistique ANOVA qui étudie la corrélation entre une variable qualitative (le type de cancer :CLINIC.study)et une variable quantitative(l'expression du gène :EXP.BRCA2) :

Entrée [77]:
```
#ANOVA (F-TEST) quanti-quali
df_anova = df[['EXP.BRCA2','CLINIC.study']]
grps = pd.unique(df['CLINIC.study'].values)
print(grps)
```

```
['prad' 'paad' 'brca']
```

Entrée [78]:
```
#ANOVA (F-TEST)
df_anova = df
df_anova = df_anova[['EXP.BRCA2','CLINIC.study']]
grps = pd.unique(df_anova['CLINIC.study'].values)
print(grps)
d_data = {grp:df_anova['EXP.BRCA2'][df_anova['CLINIC.study'] == grp] for grp in
```

```
['prad' 'paad' 'brca']
```

Entrée [79]:
```
F, p = f_oneway(d_data['prad'], d_data['paad'])
print("p-value for significance is: ", p)
if p<0.05:
    print("rejeter les hypotheses nulles")
else:
    print("accepter les hypotheses nulles")
```

```
p-value for significance is:  1.8787265068554717e-46
rejeter les hypotheses nulles
```

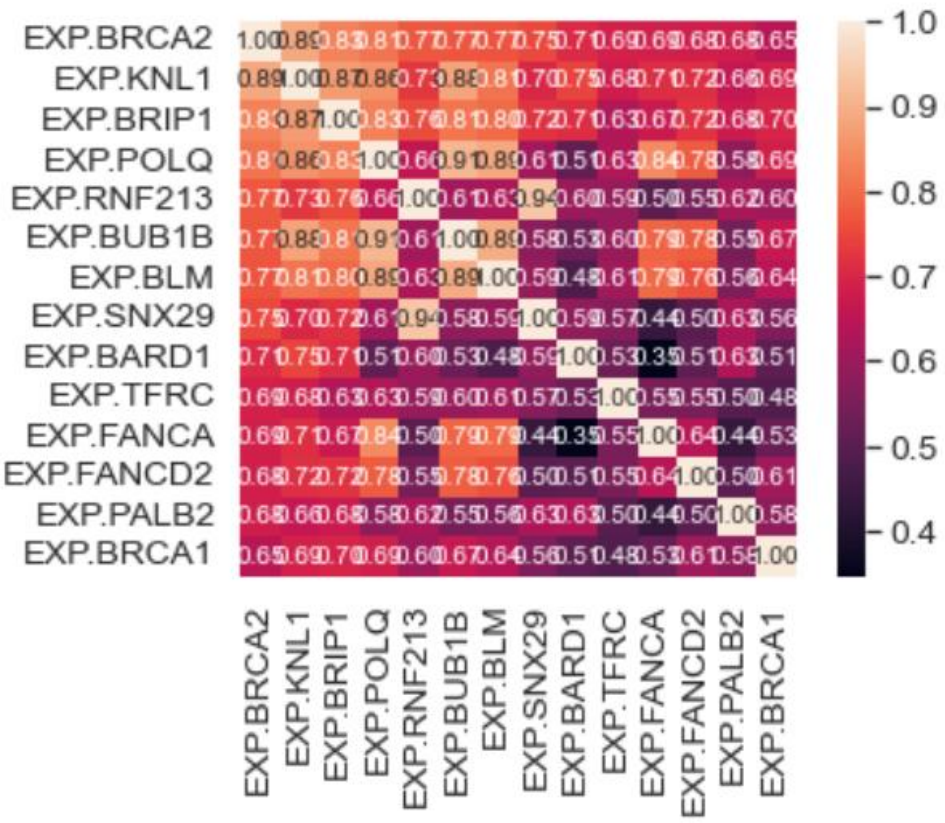
l'hypothèse nulle met en évidence l'égalité des moyennes : la variable qualitative n'a aucune influence sur la variable quantitative. enrejettnat cette hypothese on justifie la relation entre les deux variables

Nous obtenons une p-value inférieur à 0.5 ce qui nous permet de rejeter les hypothèse nulle qui dit qu'il n'y a pas de corrélation entre les deux variables. En conséquence, la corrélation est bien démontrée.

Maintenant, pour continuer notre analyse, on pourrait se demander quels sont les gènes les plus corrélés avec BRCA2 ?

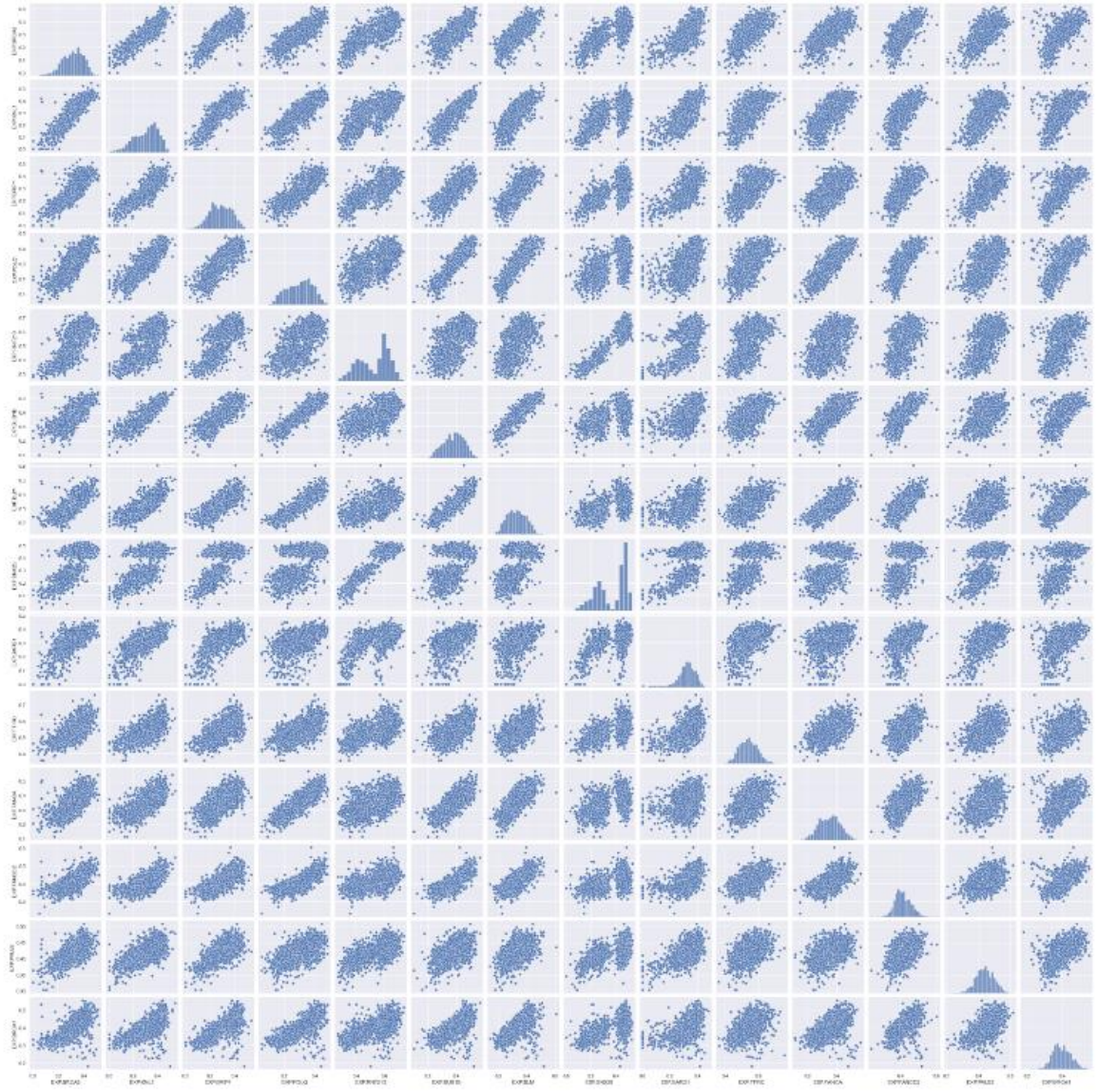
En d'autres termes, quels sont les gènes qui sont affectés par la mutation du gène clé BRCA2 dans les cas de cancers de Prostate et de Pancréas?

Pour répondre à cette question, nous allons choisir les 13 gènes les plus corrélées avec le gène BRCA2 afin de représenter leur matrice de corrélation :



On remarque une grande corrélation qui varie entre 65% et 89% entre ces 14 gènes donc on peut considérer que ces 14 gènes sont les plus affectés par la mutation du gène BRCA2.

Nous allons représenter les graphiques de corrélation correspondante :

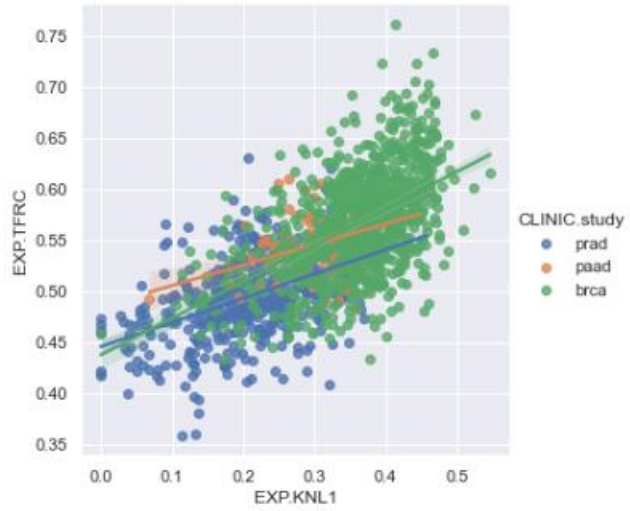


En regardant la grille des corrélations ci-dessus, nous remarquons des corrélations positives fortes (identique au Heatmap).

Voyons de plus près un figure parmi ceux-dessus :

```
Entrée [154]: sns.lmplot(x="EXP.KNL1", y="EXP.TFRC", hue="CLINIC.study", data=df)
```

```
Out[154]: <seaborn.axisgrid.FacetGrid at 0x2b253bd3640>
```



Nous remarquons bien que la corrélation est très similaire pour les trois types de cancers, et que la distribution des points du cancer de sein (brca)sont un peu plus décalés vers la droite et vers le haut. Ce qui vaut dire que dans le cas du cancer de sein l'expression de ces deux gènes est plus fort que dans le cas des cancers de prostate ou de pancréas(paad et prad).

Revenons au grille de corrélation nous remarquons aussi des nuages de points qui prennent un forme étrange, et qui forment deux nuages de points dans un même figure!

Ceci est très facilement remarquable pour les deux gènes :SNX29 ,RNF213.

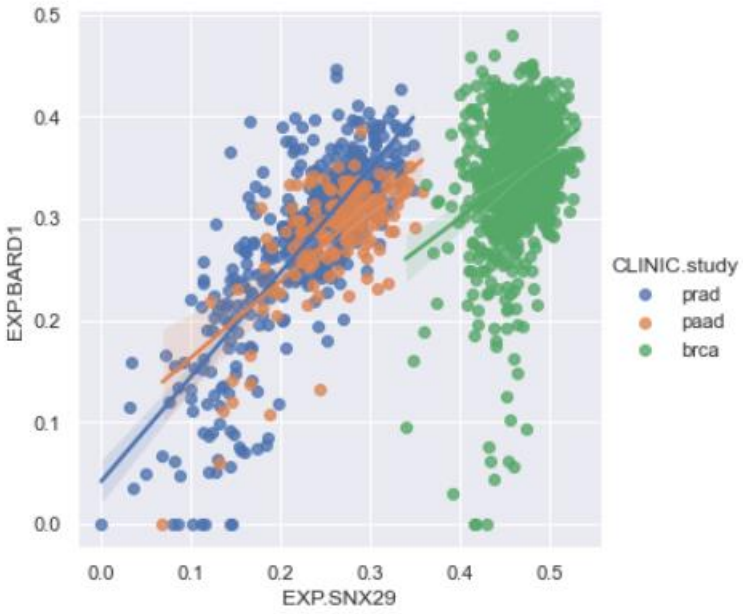
Nous pouvons donc dire que ces deux gènes ont un comportement plus fort et distinct par rapport aux autres gènes.

Nous allons se concentrer sur ces deux gènes en particulier afin de comprendre la raison de cette comportement dans ce qui suit.

Pour voir les figures de plus près, présentons la corrélation entre le 1er gène en question SNX29 et un autre gène dans la liste des 14 gènes prenons par exemple BARD1 :

```
Entrée [235]: sns.lmplot(x="EXP.SNX29", y="EXP.BARD1", hue="CLINIC.study", data=d
```

```
Out[235]: <seaborn.axisgrid.FacetGrid at 0x1f69885fa90>
```



Nous remarquons une chose importante: le nuage de points du cancer de sein est concentré le plus à droite et en haut dans le figure ce qui vaut dire que dans le cancer de sein l'expression de ces deux gènes est bien plus élevé par rapport aux cancers du Pancréas et du Prostate.

C'est-à-dire, pour le cancer du Sein l'expression du gène SNX29 est bien plus élevé (entre 0.4 et 0.5) par rapport aux deux autres type de Cancer (entre 0.1 et 0.35) ce qui forme bien deux nuages de points. Et pareil pour le gène BRAD1.

Nous pouvons remarquer que lorsque l'expression de ces deux gènes est petit cela signifie un cancer de Prostate ou de Pancréas et lorsque l'expression est grande cela signifie un cancer de Sein.

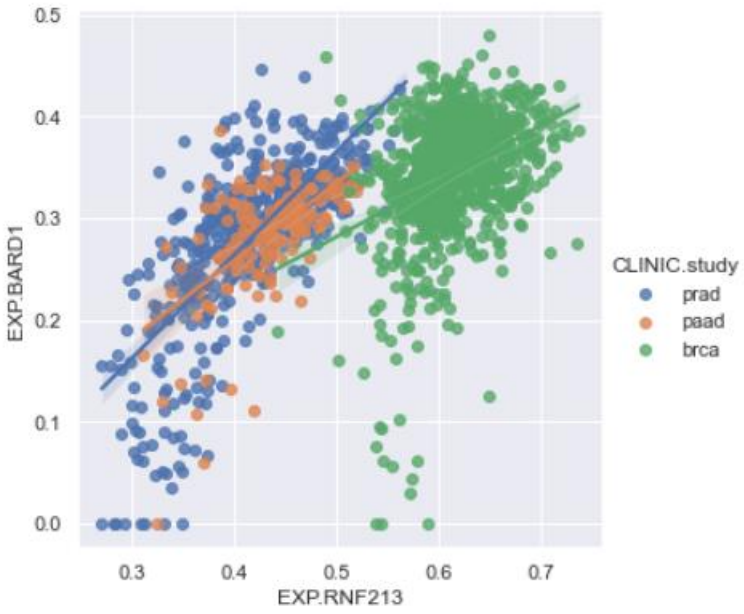
Nous remarquons aussi une corrélation positive forte entre ces deux gènes lors qu'il s'agit du cancer de pancréas ou de prostate(paad et prad), tandis que lorsqu'il s'agit du cancer de Sein il n'est plus le cas puisque les points sont distribués verticalement.

En d'autres termes, lorsque l'expression du gène SNX29 augmente, l'expression du gène BARD1 augmente aussi et vice versa cela est pour les deux cancers Prostate et Pancréas.

Alors que pour le cancer du Sein, l'augmentation de l'un n'entraine pas une augmentation (ou diminution) de l'autre, c'est-à-dire les deux gènes ne sont pas très corrélées.

Présentons maintenant la corrélation entre le second gène en question RNF213 et le gène BARD1 :

```
Entrée [242]: sns.lmplot(x="EXP.RNF213", y="EXP.BARD1", hue="CLINIC.study", data=df)
Out[242]: <seaborn.axisgrid.FacetGrid at 0x1f6a821af70>
```



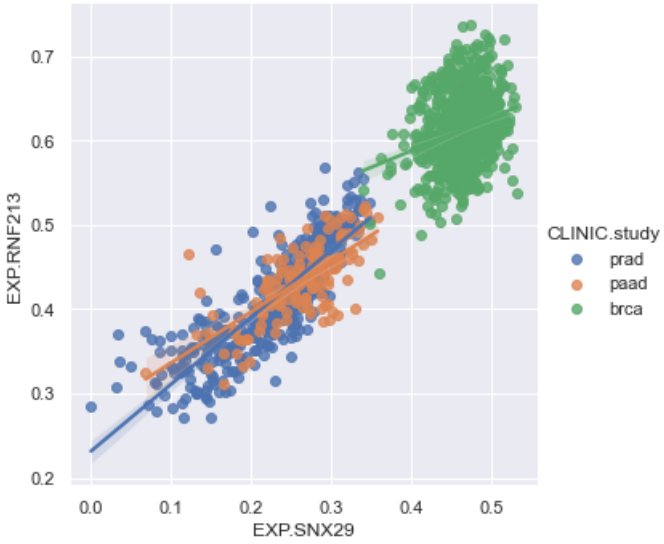
Identiquement au figure précèdent, nous remarquons aussi que le nuage de points du cancer de sein est concentré le plus à droite et en haut dans le figure ce qui vaut dire que dans le cancer de sein l'expression de ces deux gènes est bien plus élevé par rapport aux cancers du pancréas et du prostate(paad et prad),.

Une corrélation positive forte entre ces deux gènes lors qu'il s'agit du cancer de pancréas ou de prostate(paad et prad), alors que lorsqu'il s'agit du cancer de sein les points sont distribués presque verticalement.

Prenons cette fois-ci les deux gènes en question ensemble et voyons leur corrélation :

```
ıtrée [26]: sns.lmplot(x="EXP.SNX29", y="EXP.RNF213", hue="CLINIC.study", data=df)
r,p=pearsonr(df['EXP.SNX29'],df['EXP.RNF213'])
print('p-value=',p)
#autre méthode
my_rho = np.corrcoef(df['EXP.SNX29'], df['EXP.RNF213'])
print(my_rho)
```

```
p-value= 0.0
[[1.         0.93636739]
 [0.93636739 1.        ]]
```



Avec une corrélation de 93%, on remarque cette fois deux populations remarquablement très distinctes, un groupe qui contient les deux type de cancer prostate et pancréas qui sont superposées à gauche en bas et un autre groupe décalé à droite en haut pour le cancer de sein.

Donc, on peut dire que l'expression faible de ces deux gènes signifie un cancer de Prostate ou de Pancréas alors que un expression fort de ces deux gènes signifie un cancer de sein.

Nous remarquons la corrélation positive forte entre ces deux gènes lors qu'il s'agit du cancer du Pancréas et du Prostate, alors que la corrélation presque nulle pour le cancer du Sein.

Nous avons pu montrer que le gène BRCA2 affecte l'expression d'un groupe des gènes qui est corrélé avec. Puisque un expression forte de ces gènes indique un cancer de sein et un expression faible indique un cancer de Pancréas ou de Prostate.

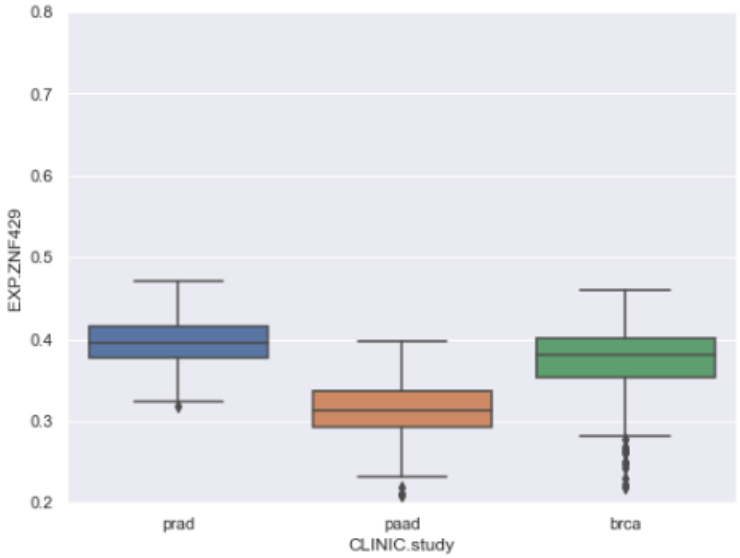
C’est-à-dire que pour les gènes appartenant aux 14 gènes, ses expressions est remarquablement plus élevé dans le cas du cancer de sein par rapport aux autres types de cancer.

Vérifions cela avec tous les 14 gènes :

C’est bien démontré !

Alors que pour un gène en dehors ces 14 gènes on trouve :

Entrée [50]:
```
var = 'CLINIC.study'
data = pd.concat([df['EXP.ZNF429'], df[var]], axis=1)
f, ax = plt.subplots(figsize=(8, 6))
fig = sns.boxplot(x=var, y="EXP.ZNF429", data=df)
fig.axis(ymin=0.2, ymax=0.8);
```



L'expression du gène ZNF429(qui n'est pas dans le groupe des 14 gènes) n'est pas remarquablement plus élevé pour le cancer de sein. Donc on garde notre résultat précédent.

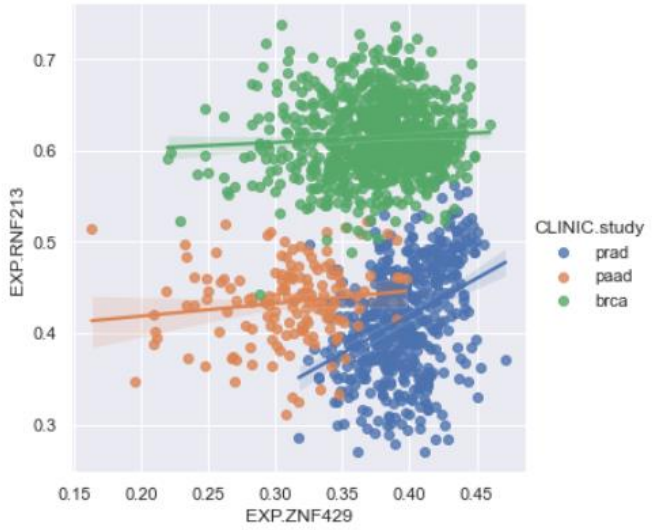
Reprenons une autre exemple, dans laquelle on va prendre le gène en question RNF213 mais cette fois-ci avec un gène en d'hors du liste des 14 gènes les plus corrélés par exemple, ZNF429.

Et avant de regarder le graphique correspondant, selon notre résultat précédent, on attend à voir :

- une très faible corrélation entre les deux gènes dans les trois types de cancers,
- ainsi que une expression plus élevé pour le gène RNF213 dans le cas du cancer de sein par rapport aux deux autres types de cancer,
- mais pour le gène ZNF429 qui n'appartient pas au group de 14 gènes son expression ne doit pas être remarquablement diffèrent selon le type de cancer!

Regardons maintenant le graphique :

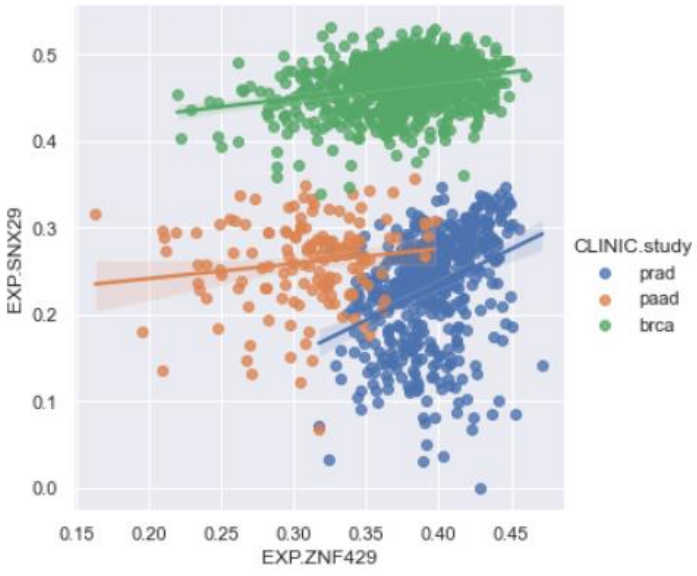
```
Entrée [22]: sns.lmplot(x="EXP.ZNF429", y="EXP.RNF213", hue="CLINIC.study", data=df)
Out[22]: <seaborn.axisgrid.FacetGrid at 0x1d227060b20>
```



Nous retrouvons exactement ce qu'on a supposé. Une faible corrélation entre les deux gènes pour les trois types de cancer, et une expression élevé pour le gène RNF213 dans le cas du cancer de sein alors que c'est moins élevé pour les cancers de prostate et de pancréas.

Pour vérifier cela, reproduisons la même expérience mais avec le gène SNX29 (un gène parmi les 14 gènes et un autre en dehors du groupe) :

```
Entrée [25]: sns.lmplot(x="EXP.ZNF429", y="EXP.SNX29", hue="CLINIC.study", data=df)
Out[25]: <seaborn.axisgrid.FacetGrid at 0x1d237d94d90>
```

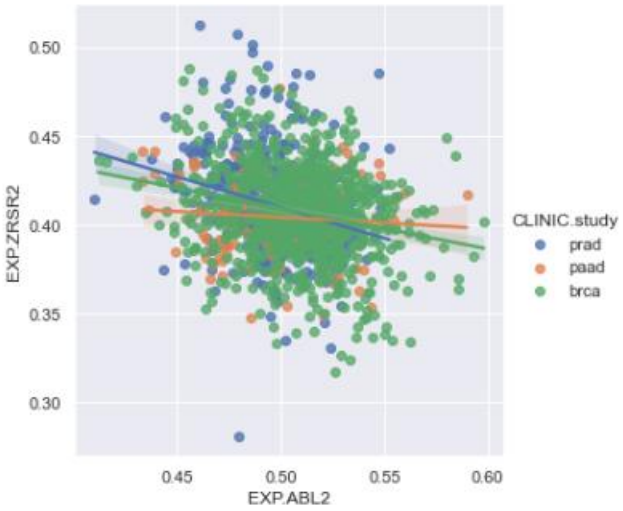


Identique au résultat précèdent ! ce qui vérifie bien notre résultat.

Prenons cet fois ci deux gènes qui n'appartiennent pas au group de 14 gènes et voyons ce que ça donne ? si nous avions raison dans le résultat précèdent, alors nous obtiendrons des nuages de points superposés et qui ne nous disent pas grandes choses. Puisque ces gènes ne seront pas responsables de cancer et par conséquent, ils ne nous donneront pas d'informations intéressants :

```
Entrée [155]: sns.lmplot(x="EXP.ABL2", y="EXP.ZRSR2", hue="CLINIC.study", data=df)
```

```
Out[155]: <seaborn.axisgrid.FacetGrid at 0x2b253bc6a30>
```



Nous obtenons exactement ce que nous avons supposé !

3.ACP :

Dans cette partie nous allons appliquer l'ACP sur les 715 gènes afin de réduire la dimensionnalité ensuite représenter les données dans des dimensions réduites pour voir si les données formeront des clusters distincts ou non.

Dans un premier temps, nous allons coder la variable de CLINIC.study en des chiffres afin de pouvoir représenter les graphique coloriés selon le type de cancer (avec un chiffre et ne pas avec un string) à l'aide de OrdinalEncoder :

ACP

Entrée [74]:
```
from sklearn.preprocessing import OrdinalEncoder

ord_enc = OrdinalEncoder()
df["CLINIC.study"] = ord_enc.fit_transform(df[["CLINIC.study"]])
df["CLINIC.study"]
```

Out[74]:
```
1       2.0
2       2.0
3       2.0
4       2.0
5       2.0
       ...
1452    2.0
1453    2.0
1454    2.0
1455    2.0
1456    2.0
Name: CLINIC.study, Length: 1456, dtype: float64
```

Entrée [75]:
```
y=df['CLINIC.study']
```

Avec 0 le cancer de Sein et 1 le cancer de Pancréas et 2 le cancer de Prostate

Dans un deuxième temps, nous choisissons une nombre de dimension qui conserve 76% de la variance des données.

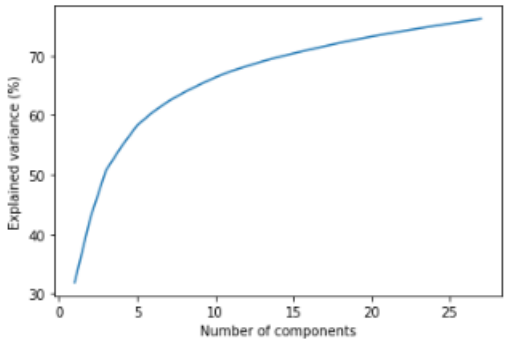
Avec un tel variance le modèle de PCA me propose de conserver 27 dimensions.

Nous représentons la variance cumulée pour chaque dimension et le graphique de variance correspondant :

Entrée [80]:
```
from sklearn.decomposition import PCA
pca = PCA(n_components = 0.76)
pca.fit(X)
print("Cumulative Variances (Percentage):")
print(np.cumsum(pca.explained_variance_ratio_ * 100))
components = len(pca.explained_variance_ratio_)
print(f'Number of components: {components}')
# Make the scree plot
plt.plot(range(1, components + 1), np.cumsum(pca.explained_variance_ratio_ * 100))
plt.xlabel("Number of components")
plt.ylabel("Explained variance (%)")
```

```
Cumulative Variances (Percentage):
[31.889892   42.70515154 50.70860781 54.75198258 58.29876804 60.51973682
 62.35101433 63.83345305 65.14940966 66.35697954 67.40177628 68.22476228
 69.04033927 69.7393076  70.39139901 70.99661978 71.59717082 72.1761797
 72.71888877 73.2233033  73.68685622 74.14527579 74.57391037 74.99753295
 75.40039419 75.79665889 76.17874226]
Number of components: 27
```

Out[80]:
```
Text(0, 0.5, 'Explained variance (%)')
```



Nous allons maintenant créer une méthode qui nous affiche pour chaque dimension de l'ACP les deux gènes les plus importants :

Entrée [82]:
```
print('Top 2 most important features in each component')
print('==============================================')
for row in range(pca_components.shape[0]):
    # get the indices of the top 4 values in each row
    temp = np.argpartition(-(pca_components[row]), 2)

    # sort the indices in descending order
    indices = temp[np.argsort((-pca_components[row])[temp])][:2]

    # print the top 2 feature names
    print(f'Component {row}: {df.columns[indices].to_list()}')
```

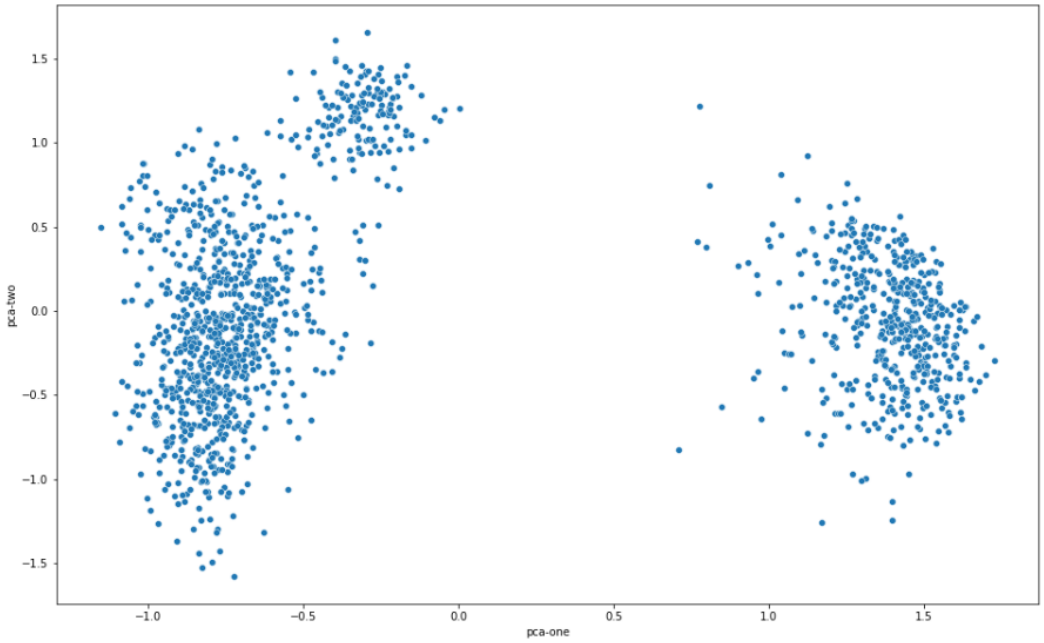
```
Top 2 most important features in each component
==============================================
Component 0: ['EXP.KLF6', 'EXP.SLC34A2']
Component 1: ['EXP.ERG', 'EXP.FNBP1']
Component 2: ['EXP.IL21R', 'EXP.APOBEC3B']
Component 3: ['EXP.ERG', 'EXP.RUNX1T1']
Component 4: ['EXP.CD74', 'EXP.TCF7L2']
Component 5: ['EXP.WDCP', 'EXP.SKI']
Component 6: ['EXP.COL1A1', 'EXP.RUNX1T1']
Component 7: ['EXP.COL1A1', 'EXP.SKI']
Component 8: ['EXP.MUC16', 'EXP.BLM']
Component 9: ['EXP.COL1A1', 'EXP.HOXA9']
Component 10: ['EXP.HOXA9', 'EXP.PTPRK']
Component 11: ['EXP.RUNX1T1', 'EXP.COL1A1']
Component 12: ['EXP.COL1A1', 'EXP.RNF43']
Component 13: ['EXP.RUNX1T1', 'EXP.CNOT3']
Component 14: ['EXP.SKI', 'EXP.CNOT3']
Component 15: ['EXP.REL', 'EXP.MUC1']
Component 16: ['EXP.FAM131B', 'EXP.RUNX1T1']
Component 17: ['EXP.CNOT3', 'EXP.COL1A1']
Component 18: ['EXP.WRN', 'EXP.BLM']
Component 19: ['EXP.NPM1', 'EXP.SKI']
Component 20: ['EXP.SKI', 'EXP.RNF43']
Component 21: ['EXP.MUC1', 'EXP.CNOT3']
Component 22: ['EXP.LRIG3', 'EXP.PAX5']
Component 23: ['EXP.LRIG3', 'EXP.CTNND1']
Component 24: ['EXP.WDCP', 'EXP.BLM']
Component 25: ['EXP.RNF43', 'EXP.CNOT3']
Component 26: ['EXP.MUC1', 'EXP.SOCS1']
```

On remarque la non présence des gènes importantes mentionnés précédemment

Et affichant les deux premiers dimensions on remarque que les données sont bien regroupés en trois clusters distincts :

```
Entrée [42]: plt.figure(figsize=(16,10))
             sns.scatterplot(
                 x=df['pca-one'], y=df['pca-two'],
                 data=X_pca,
                 legend="full",
             )
```

```
Out[42]: <AxesSubplot:xlabel='pca-one', ylabel='pca-two'>
```

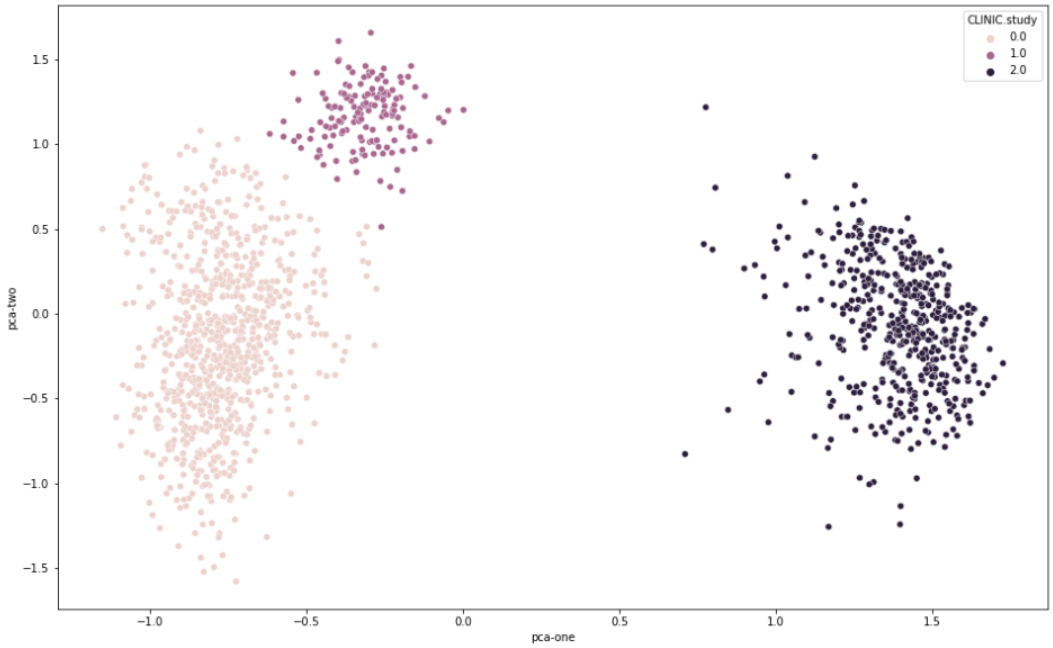


Nous remarquons bien la présence de trois clusters de données, mais de quoi s'agissent-ils ?

Colorions les données selon le type de cancer :

Entrée [87]:
```
plt.figure(figsize=(16,10))
sns.scatterplot(
    x=df['pca-one'], y=df['pca-two'],
    hue=df['CLINIC.study'],
    data=X_pca,
    legend="full",
)
```

Out[87]: <AxesSubplot:xlabel='pca-one', ylabel='pca-two'>



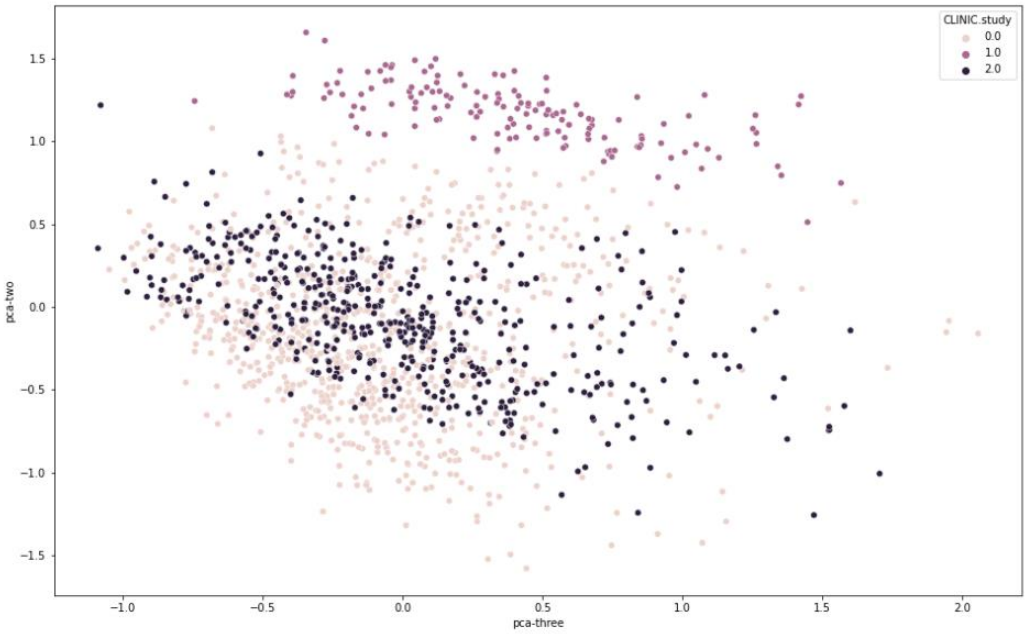
Nous pouvons remarquer que chaque type de cancer forme un cluster et que chaque patient est dans une cluster bien définit et distinct !

Nous pouvons remarquer aussi si un nouveau patient est ajouté, il aura un position précise dans l'un de ces trois clusters.

Maintenant représentons les deux dimensions 2 et 3 :

```
Entrée [88]: plt.figure(figsize=(16,10))
sns.scatterplot(
    x=df['pca-three'], y=df['pca-two'],
    hue=df['CLINIC.study'],
    data=X_pca,
    legend="full",
)
```

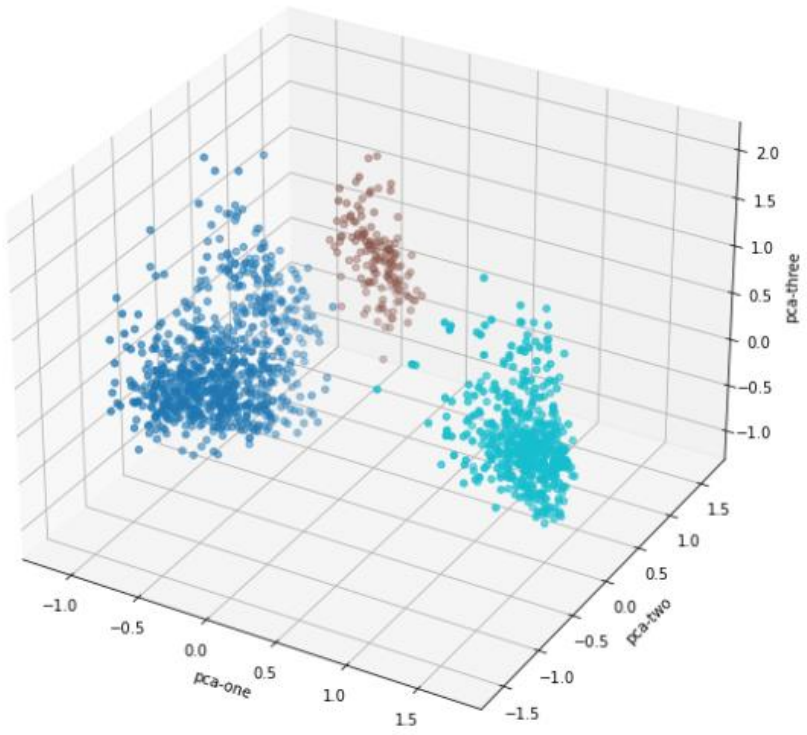
```
Out[88]: <AxesSubplot:xlabel='pca-three', ylabel='pca-two'>
```



Ici nous remarquons que, les deux types de cancer Sein et Prostate sont superposées, alors que le cancer de Pancréas est un petit peu plus loin. Ceci revient au fait de regarder les données d'une autre coté(dimension).

Et si l'on affiche les trois dimensionnalités ensemble :

Entrée [89]:
```
ax = plt.figure(figsize=(16,10)).gca(projection='3d')
ax.scatter(
    zs=df['pca-three'] ,
    xs=df['pca-one'],
    ys=df['pca-two'] ,
    c=df['CLINIC.study'],
    cmap='tab10'
)
ax.set_xlabel('pca-one')
ax.set_ylabel('pca-two')
ax.set_zlabel('pca-three')
plt.show()
```



Nous pouvons remarquer que chaque type de cancer forme un nuage de points diffèrent, ceci nous informe qu'il y a bien des variables cachés (gènes)derrière et qui pourraient nous indiquer précisément quel est le type de cancer atteint.

C'est-à-dire à partir des expressions de certains gènes, on pourra connaitre le type de cancer atteint ou va atteindre la personne prochainement.

Nous avons découvert pendant ce projet un groupe de ces variables et peut-être les professionnels peuvent nous en dire plus et nous donner une autre piste de recherche plus précise.

4.Conclusion :

Nous avons pu dans ce projet, en se basant sur un article qui nous a mis sur le premier piste de recherche, de découvrir un groupe des gènes qui pourrait être intéressant dans l'étude de cancer.

Cet article nous dit, que les mutations du gène BRCA2 peuvent entrainer des changement dans l'expression de plusieurs autres gènes qui va en conséquence augmenter le risque d'atteindre le cancer de la prostate ou de la pancréas.

Avec une analyse successive, nous avons trouvé des résultats pertinentes qui sont les suivantes :

- Nous avons recherché les 14 gènes les plus corrélés avec BRCA2 parmi les 715 gènes disponibles.
- Ces 14 gènes se comportent différemment par rapport aux autres gènes.

  C'est-à-dire, lorsque le cancer atteint est le cancer du sein, nous remarquons que l'expression de ces gènes est bien élevé. Alor que lorsqu'il s'agit du cancer de la prostate ou de la pancréas, l'expression des ces gènes est plus faible.

- Les autres gènes qui n'appartient pas à ce groupe, ne comportent pas de cette façon et ne nous permet pas d'en tirer de l'information !
- Ensuite nous avons appliqué l'ACP sur les 715 gènes et nous avons remarqué l'existence de trois clusters distincts et qui pourra nous indiquer de l'existence des gènes parlants et responsables de chaque type de cancer !

Donc peut-être s'approcher des professionnels et faire une analyse plus profonde nous permettra de tirer des informations plus précises.