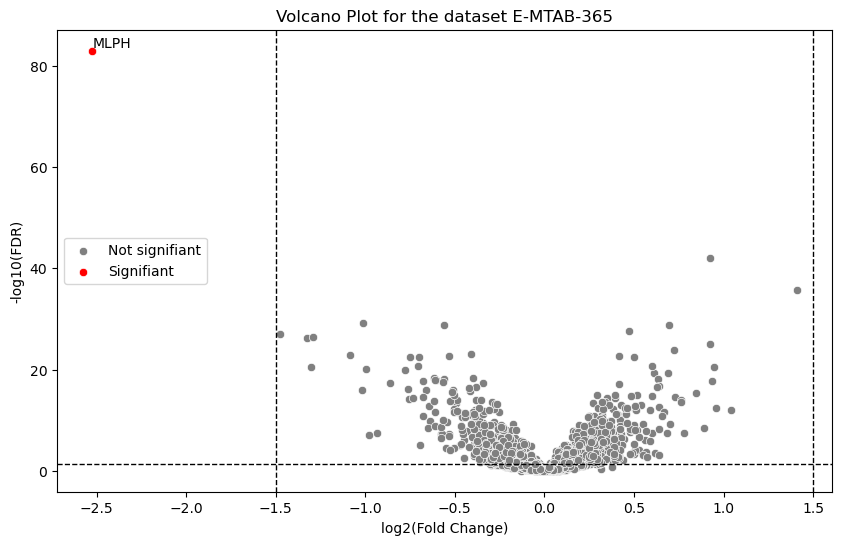
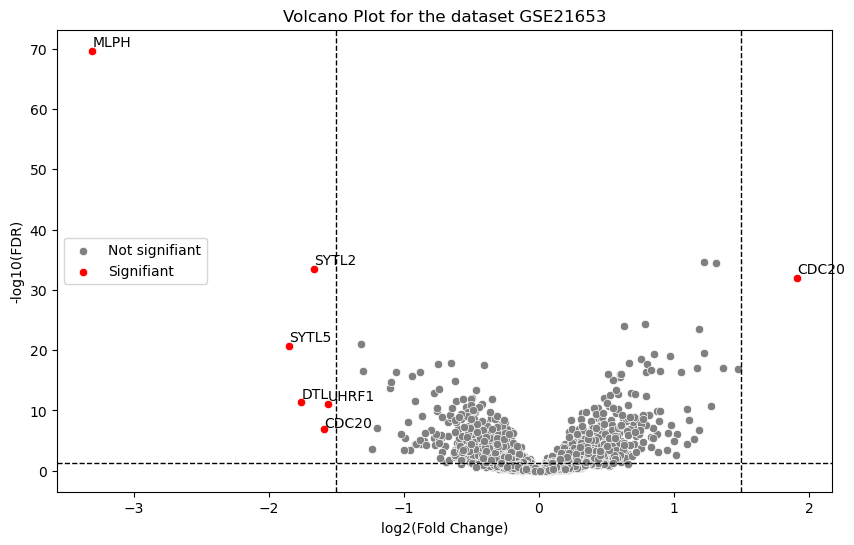
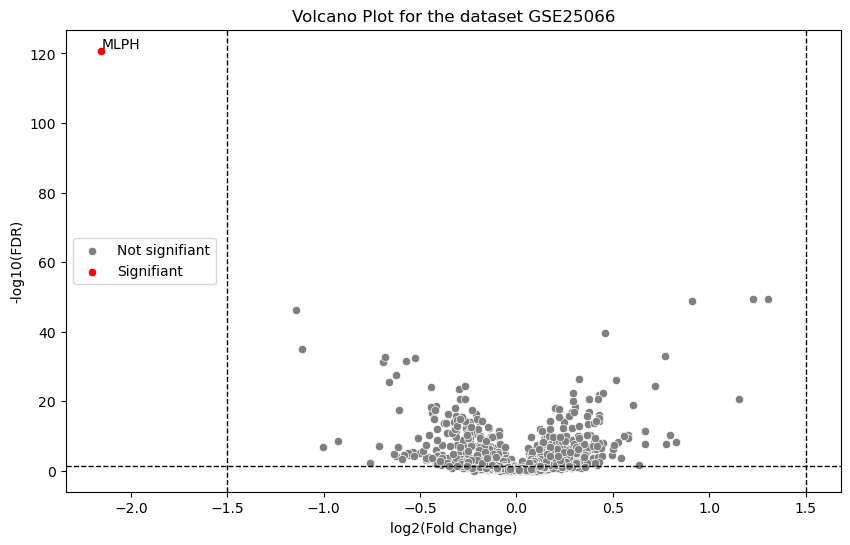
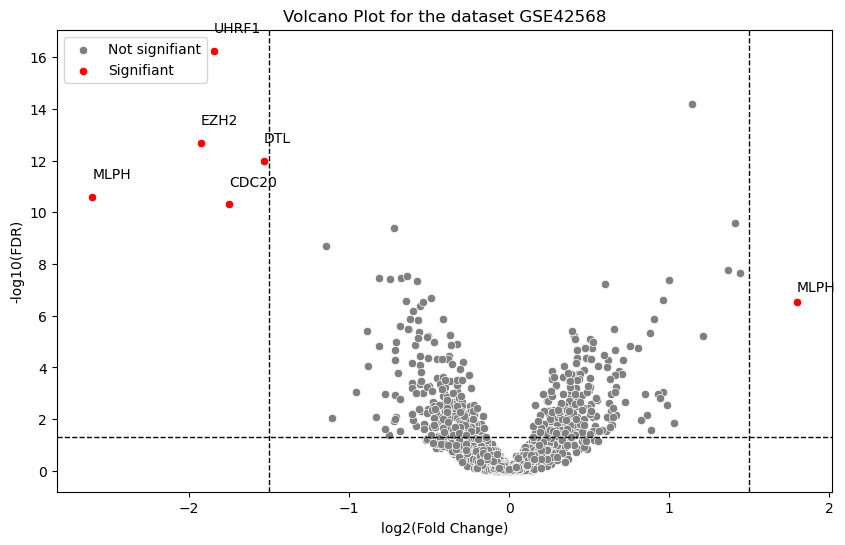
**Réalisation de Volcano plot**

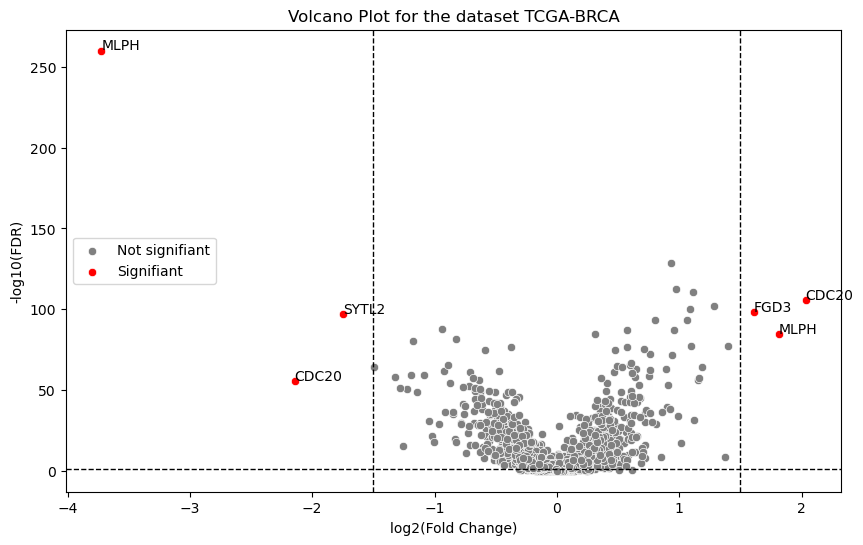
En complément de l’analyse d’expression des gènes, nous avons réalisé des Volcano plot pour chaque dataset. Cependant, dans les datasets Naderi-Caldas-2007 et Miller-2005, l’ensemble des échantillons sont du type « Non-tumours », ainsi avec de telles données, nous ne pouvons pas réaliser de volcano plot (calcul de Fold Change impossible).

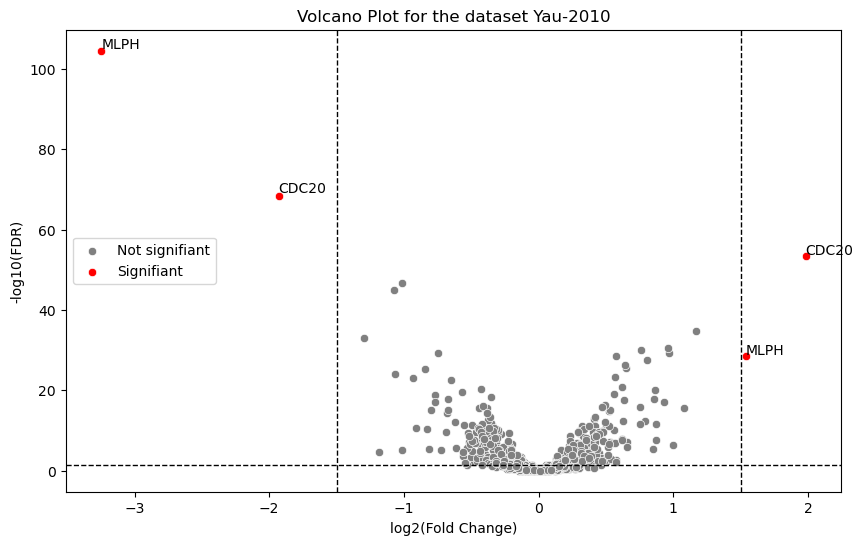
Voici les volcano plot pour l’ensemble des autres datasets :











Dans ces graphiques ci-dessus, nous représentons en rouge les points significatifs et en gris les points non significatifs. Une ligne horizontale en pointillés (qui représente les pvaleurs ajustées égales à 5%) ainsi que 2 lignes verticales en pointillées viennent séparer les points significatifs et non significatifs. Enfin, nous affichons le nom du gène sur le graphique uniquement pour les points que nous considérons comme significatifs (pour éviter de surcharger le graphique).

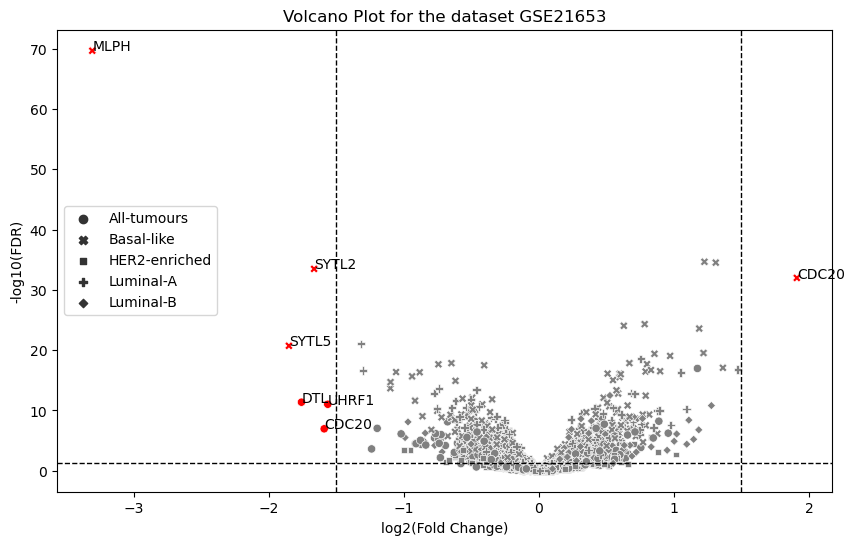
**Remarques intéressantes :**

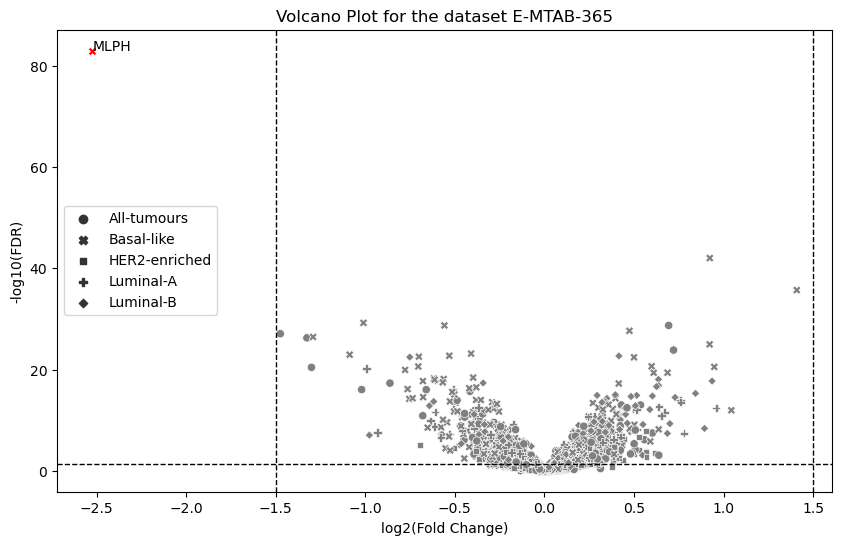
- Le gène MLPH se retrouve à la marge dans plusieurs volcano plot, sur les données GSE21653, E-MTAB-365, GSE25066, TCGA-BRCA et Yau-2010 (peut-être données abérrantes ?)

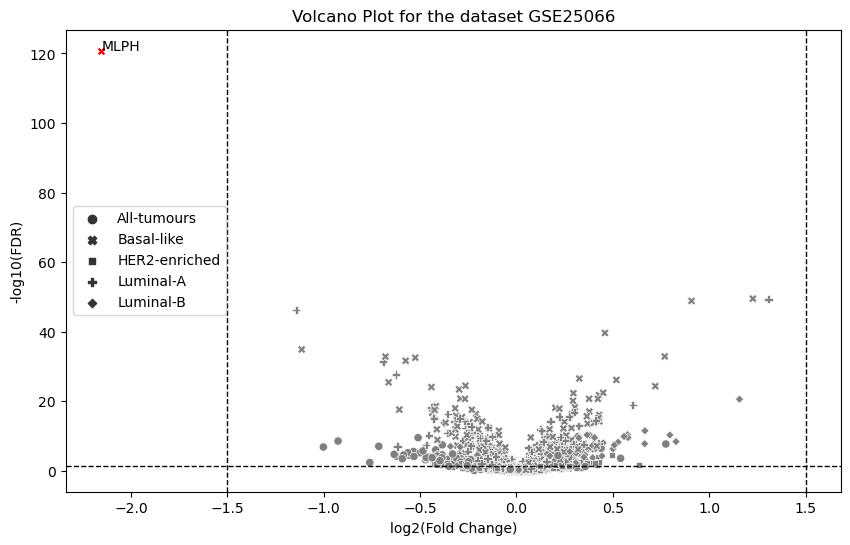
- D'un dataset à l'autre, les ordres de grandeurs sont vraiment différents. Par exemple dans TCGA des valeurs du Log du Fold Change peuvent atteindre 100 voire 200 tandis que dans GSE42568, aucun point ne dépasse la valeur 20.

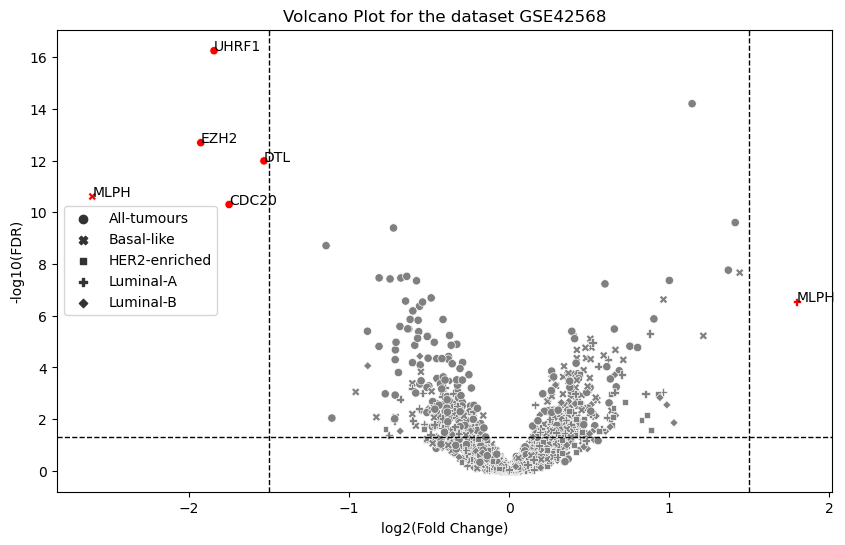
**Une variante de volcano plot avec le détail par sous types moléculaires :**

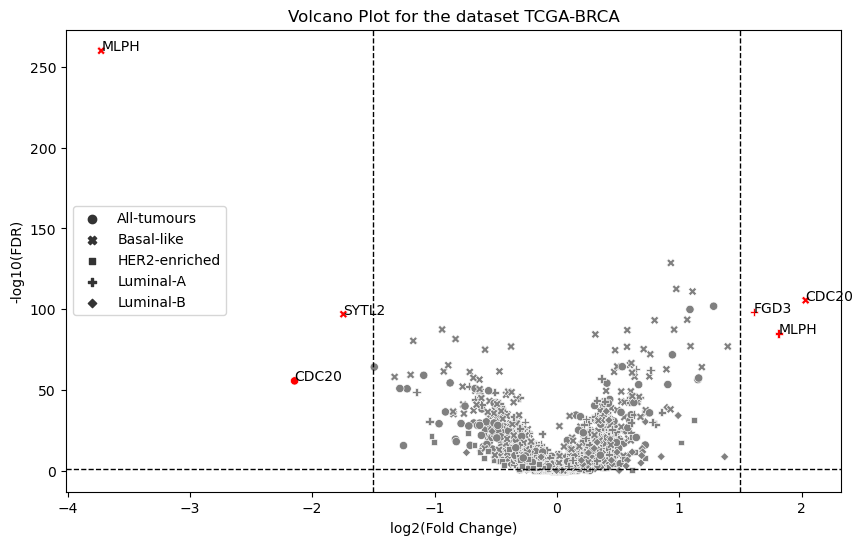
Nous avons également produit des volcano plot qui précisent le sous-type de chaque point (chaque sous-type est associé à une forme sur le graphique). Voici le résultat pour chaque sous type :

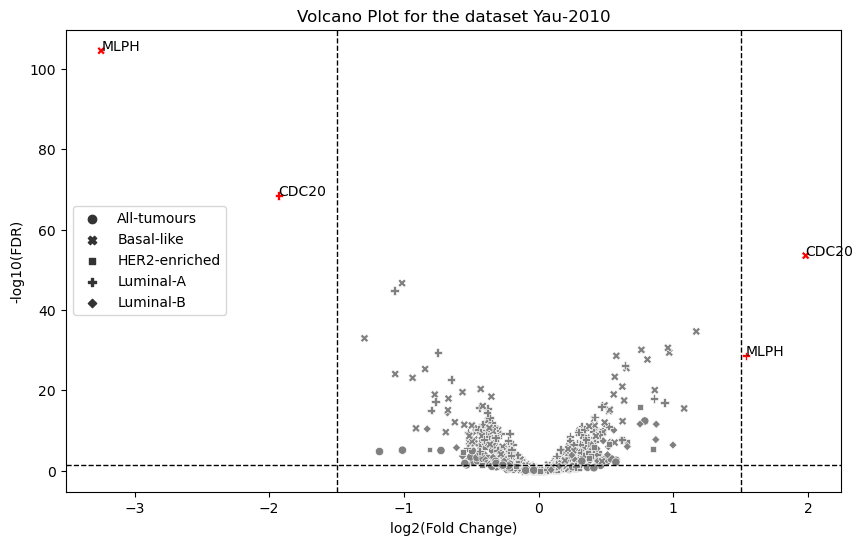


****

****



****

****

**Remarques :**

- Les points les plus significatifs sont dans chaque datasets pour le basal-like contre le reste

- On retrouve aussi beaucoup de Luminal-A

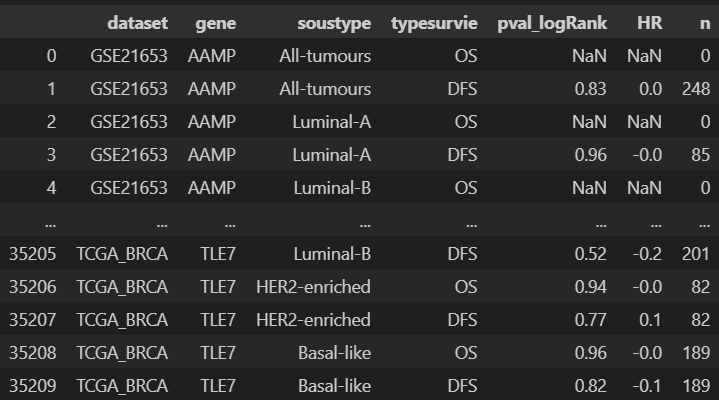
- Les autres types semblent se fondre dans la masse dans tous les datasets

**Analyse de survie**

Comme pour la partie précédente sur l’analyse d’expression des gènes, nous réalisons nos analyses sur la base réduites de 605 gènes. A noter que dans certains datasets, nous ne retrouvons pas les 605 gènes mais beaucoup moins.

**1 – Construction d’un tableau regroupant toutes les informations nécessaires**

Dans un premier temps, nous commençons par construire un tableau qui rassemble toutes les informations nécessaires pour réaliser des analyses de survie sur nos datasets. Voici un extrait de ce tableau :

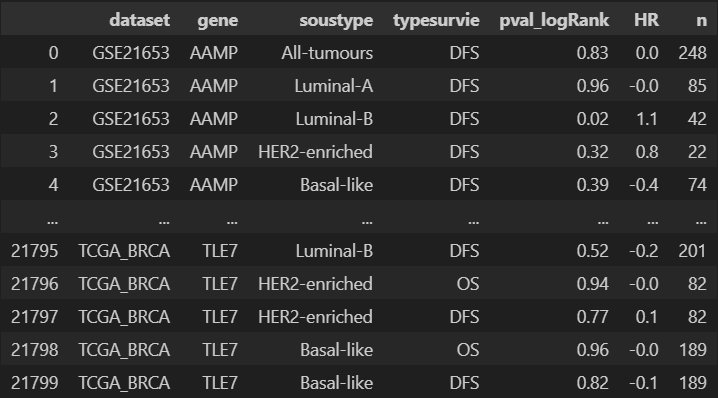
****

Ce tableau contient plus de 35000 lignes. Pour chaque ligne, nous avons des informations sur un dataset donné, un gène donné, un sous-type moléculaire donné ainsi qu’un type de survie donné. Chaque ligne contient 3 indicateurs qui permettent d’évaluer la qualité de l’analyse de survie :

* La pvaleur du test du LogRank
* Le hazard ratio
* Le nombre d’observations

Ce tableau va nous permettre de savoir quelles sont les analyses de survie les plus intéressantes à représenter graphiquement. En effet, nous ne pouvons pas représenter toutes les courbes de survies.

Pour choisir les courbes de survies à représenter, nous commençons par nous appuyer sur le nombre d’observations. Dans un premier temps, nous pouvons supprimer les lignes pour lesquelles nous n’avons aucune observation.



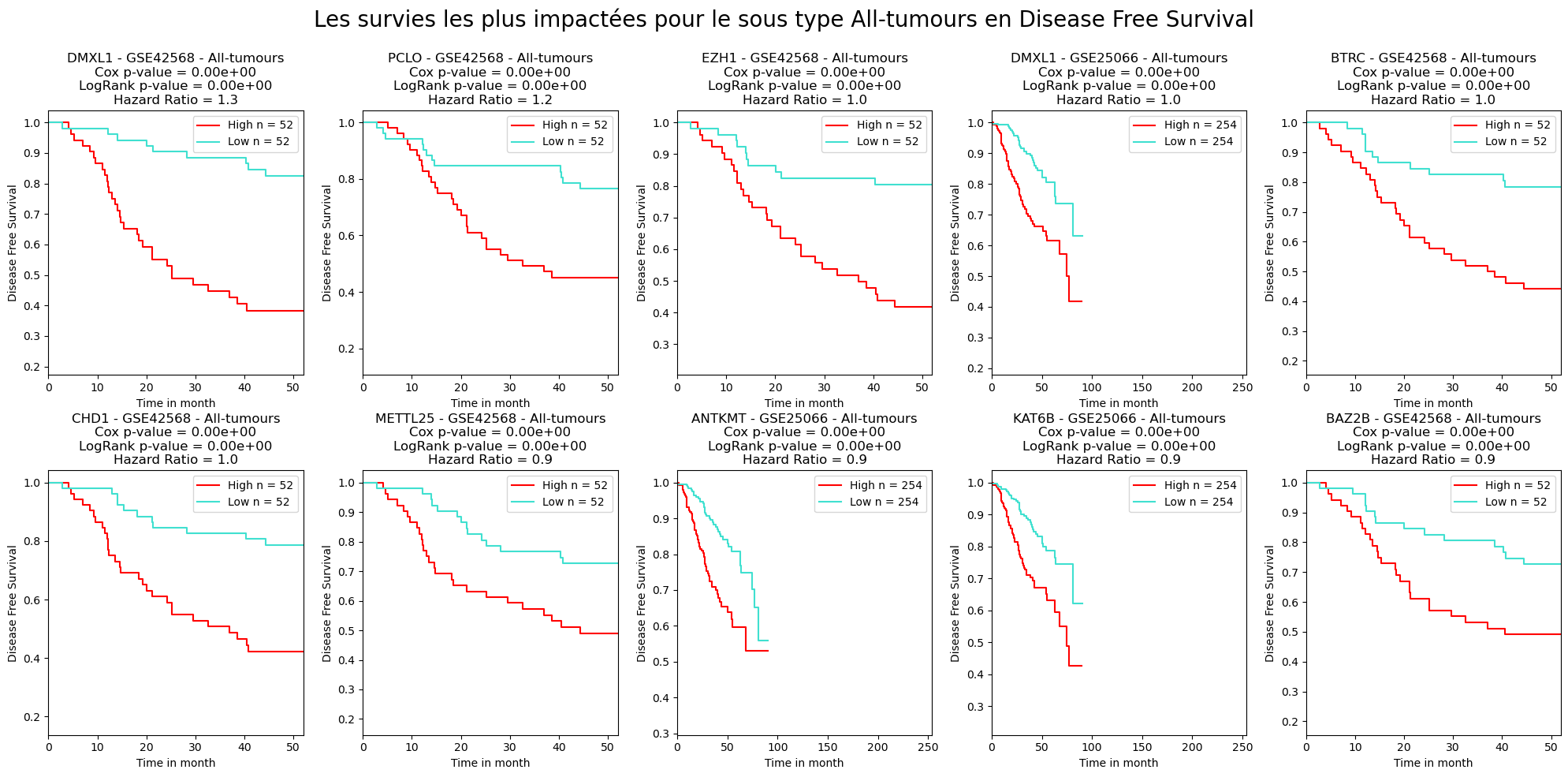
Le tableau ne contient maintenant plus qu’environ 22000 lignes.

**2 – Représenter les courbes de survies**

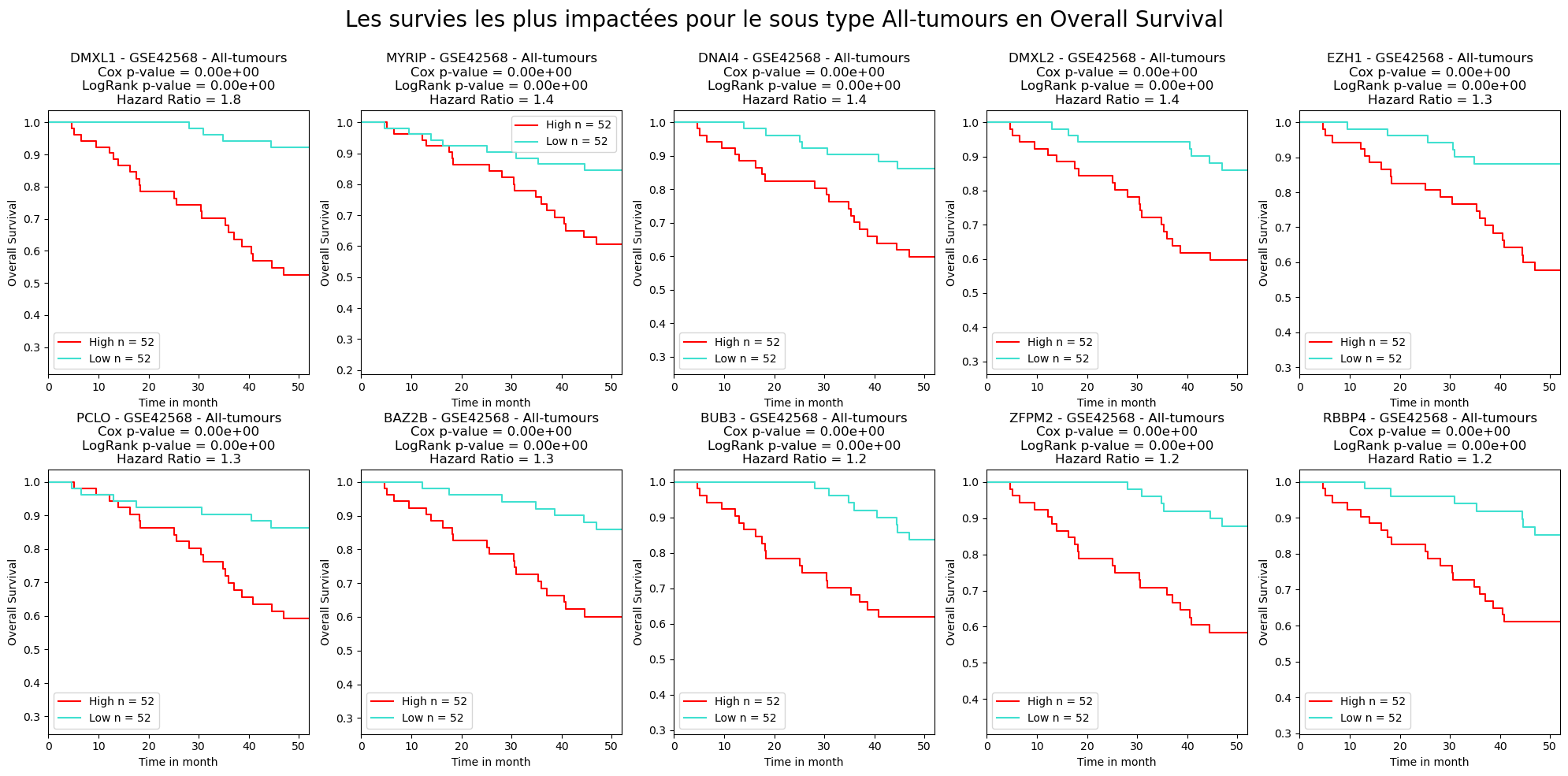
Nous allons dès à présent représenter les courbes de survies pour chaque couple : sous-type moléculaire et type de survie. Nous représentons les 2 survies les plus impactées en se basant sur les plus petites pvaleurs (du test du logrank) possible et les Hazard ratio les plus élevés possible. De plus, nous ne conserverons uniquement les courbes de survie qui contiennent au moins 100 observations.

Le sous type « **All-tumours** » :

Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « All-tumours » et pour la survie sans rechute :

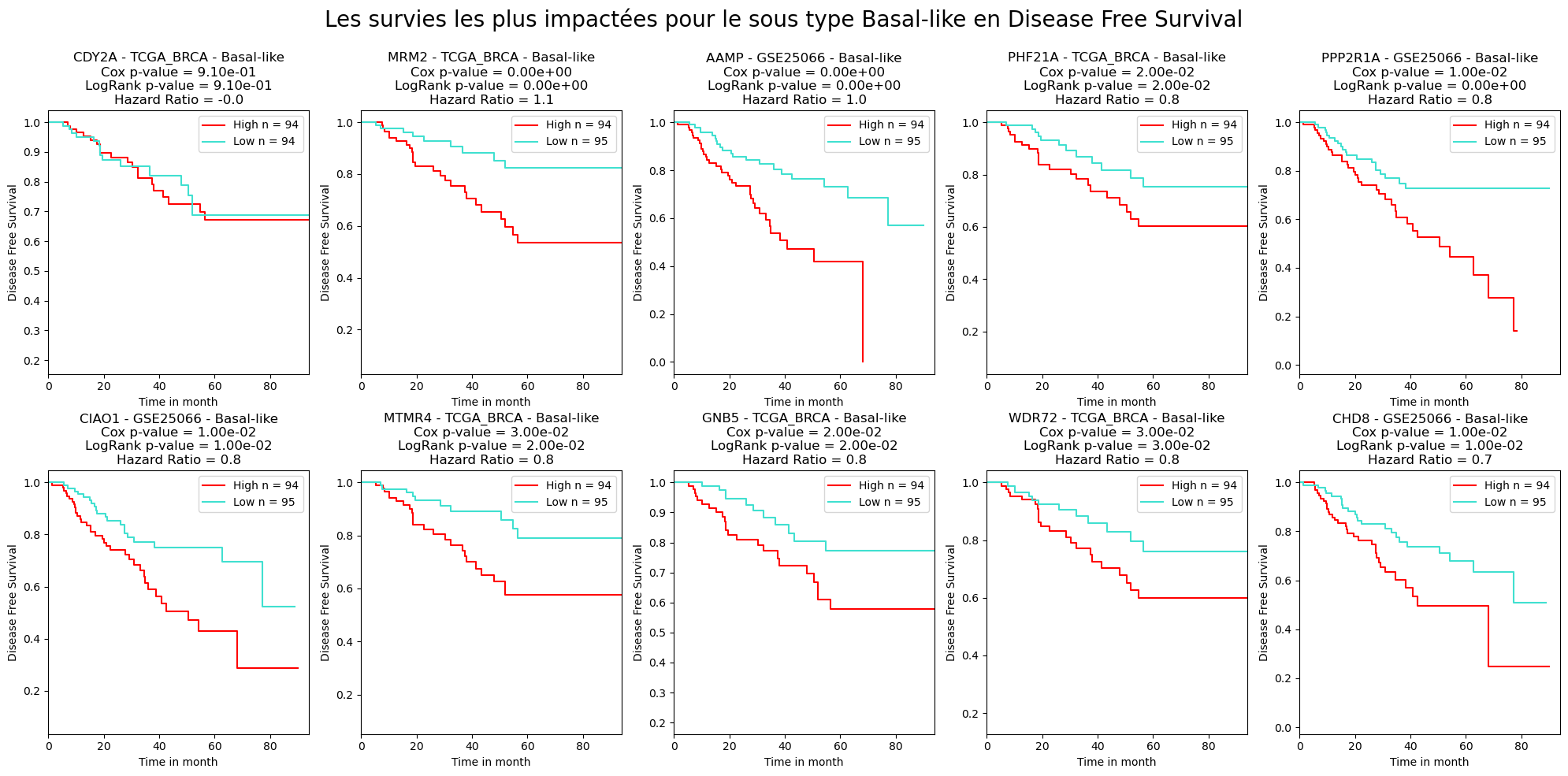


Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « All-tumours » et pour la survie globale :

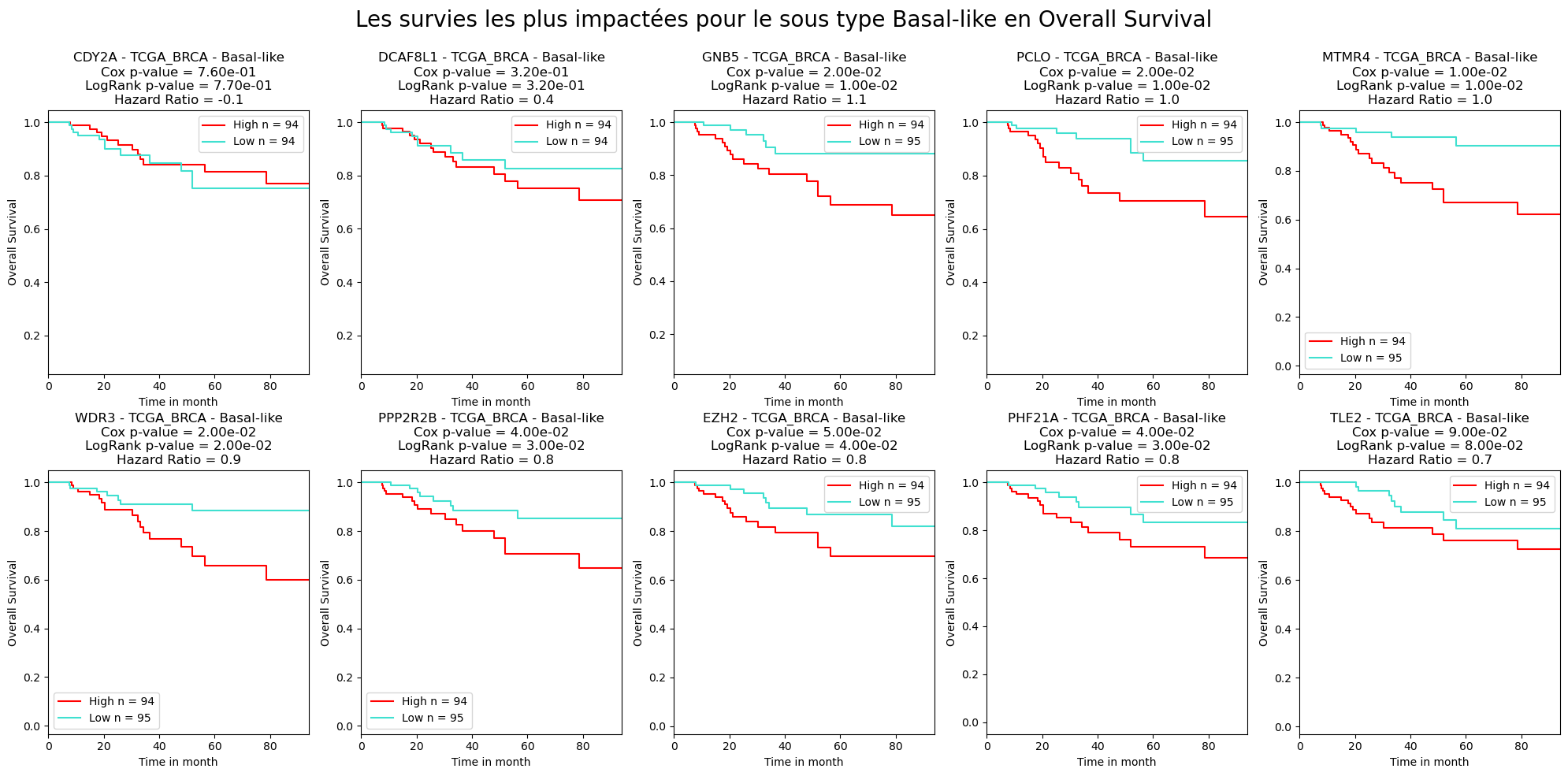


Le sous type « **Basal-like** » :

Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « Basal-like » et pour la survie sans rechute :

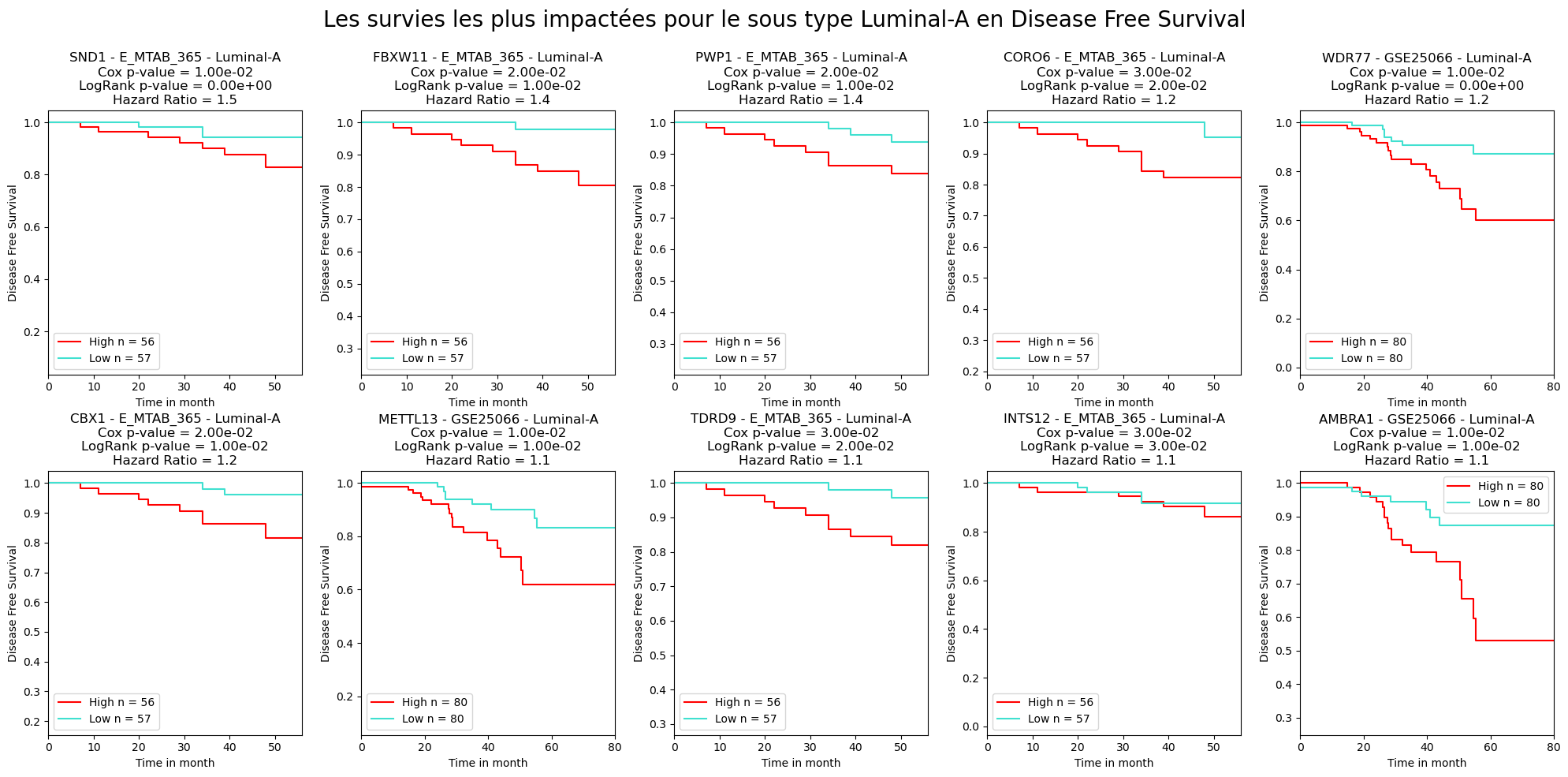


Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « Basal-like » et pour la survie globale :

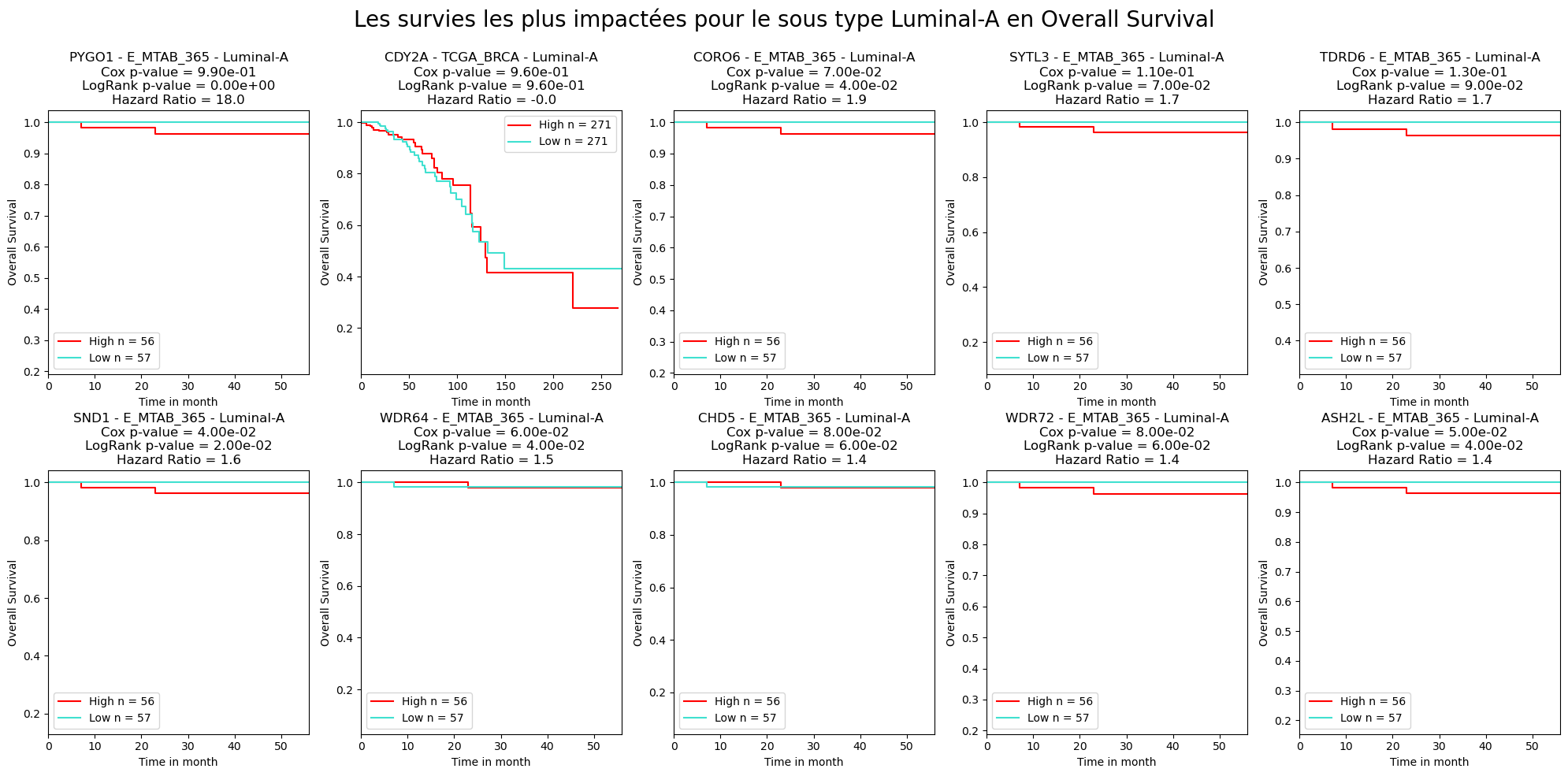


Le sous type « **Luminal-A** » :

Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « Luminal-A » et pour la survie sans rechute :

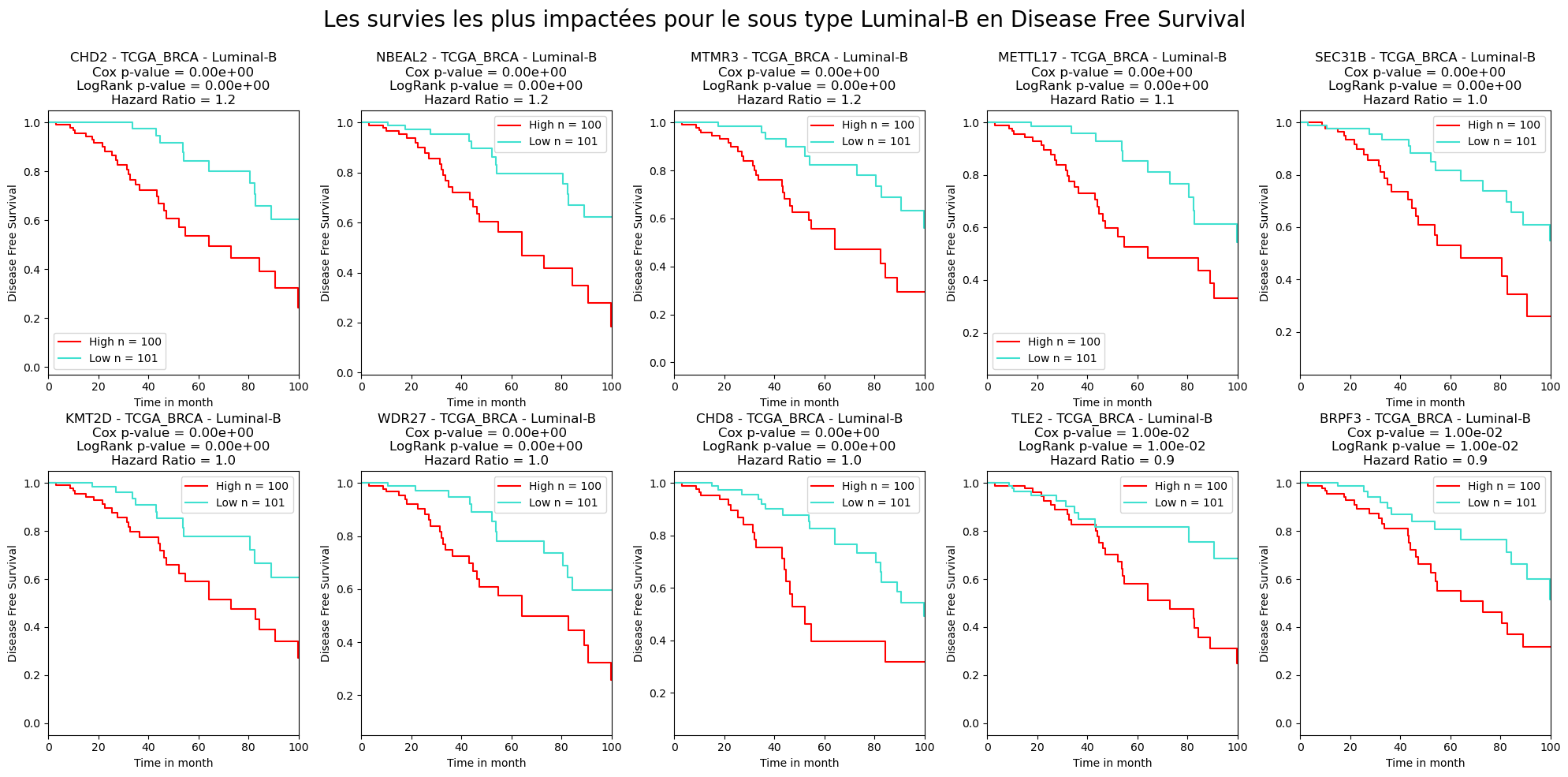


Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « Luminal-A » et pour la survie globale :

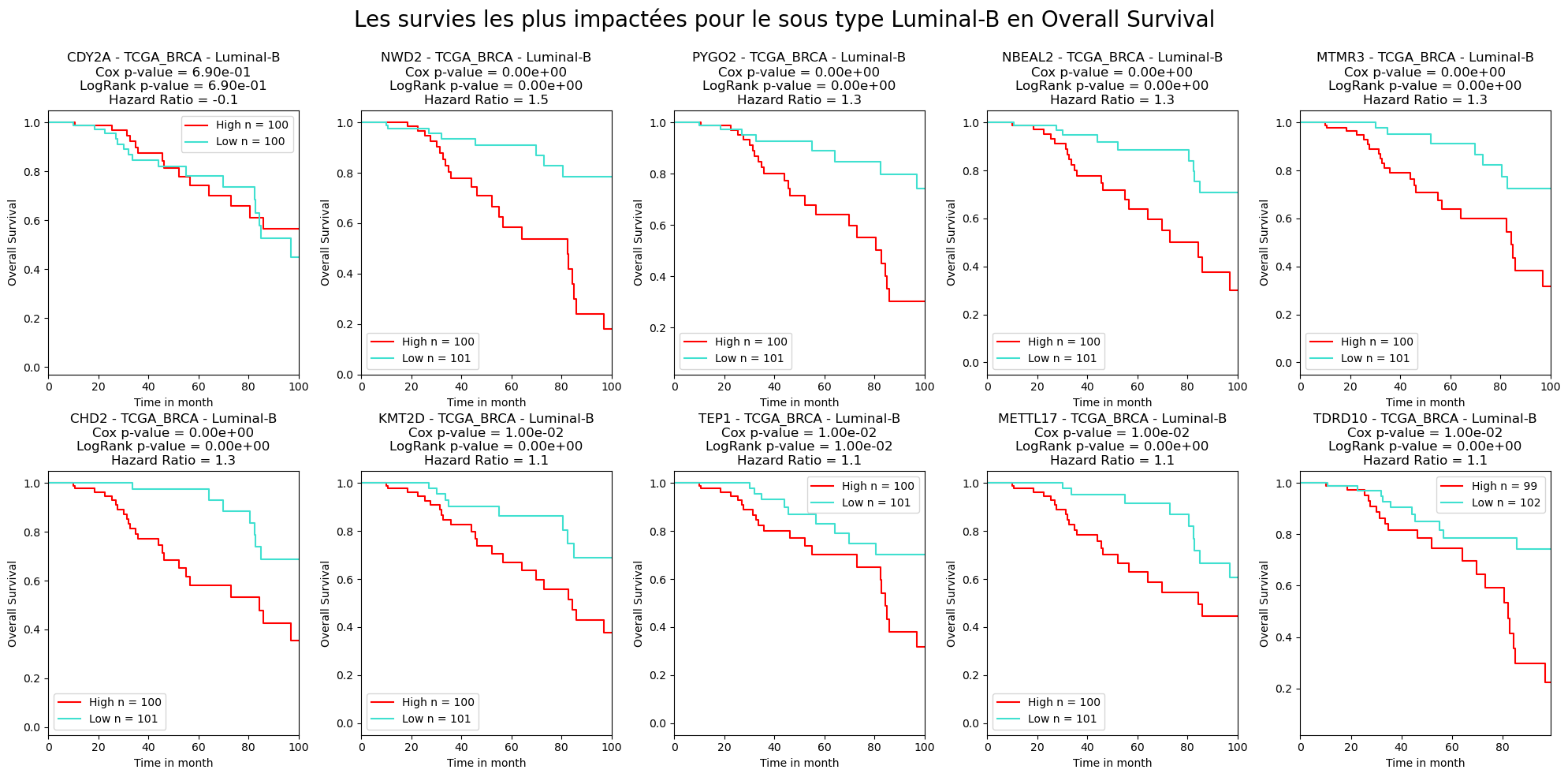


Le sous type « **Luminal-B** » :

Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « Luminal-B » et pour la survie sans rechute :



Dans le visuel ci-dessous, nous représentons les courbes où la survie est le plus impactées pour le sous type « Luminal-B » et pour la survie globale :



A noter que les données de survie pour le sous type « HER2-enriched » ne respecte pas les critères que nous avons défini (avoir un nombre d’observations supérieur à 100). Nous avons donc décider de ne pas afficher les résultats.

**3 – Identification des gènes qui impactent le plus la survie**

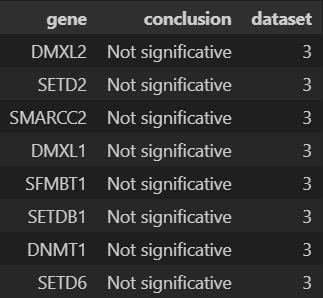
Dans cette dernière partie pour l’analyse de survie, nous allons chercher à répondre à cette question : *Quels sont les gènes dont le niveau d’expression est associé à la survie de façon robuste, c’est-à-dire, avec validation dans plusieurs datasets ?*

La réponse a cette question peut permettre d’identifier une base réduite de gènes dont les conclusions peuvent être prises au sérieux car des conclusions similaires se répètent dans plusieurs jeux de données.

Pour répondre à cette question, nous pouvons extraire un tableau qui donne la répétition d’une même conclusion pour un gène dans plusieurs jeux de données. Nous considérons 2 conclusions possibles :

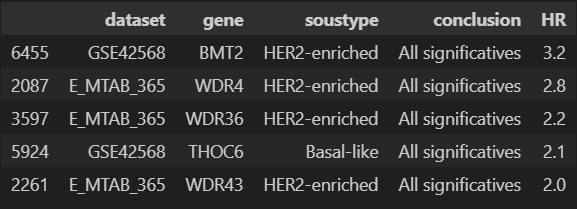
* « All significatives » : Lorsque le FDR du Logrank et du modèle de Cox sont tout 2 inférieurs à 5%.
* « Not significatives » : Lorsque qu’au moins une des 2 pvaleurs ajustées (FDR) est supérieure à 5%.

Voici un extrait des tableaux obtenus :

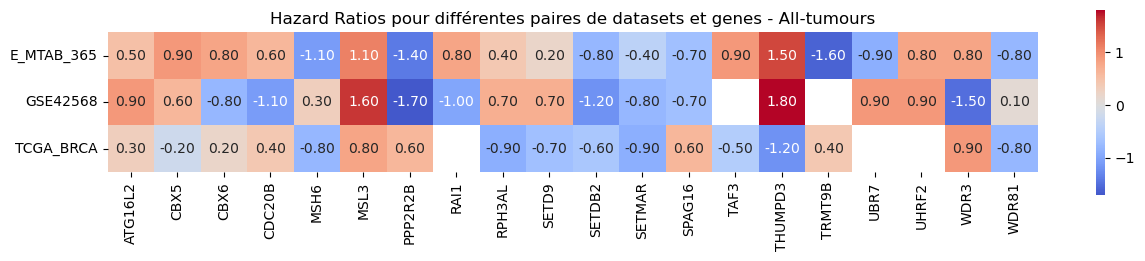


A gauche, nous avons dans l’ordre décroissant, les gènes dont la conclusion « All significatives » se répète dans le plus grand nombre de jeux de données. A droite, nous avons à l’inverse, les gènes dont la conclusion « Not significative » se répète dans le plus grand nombre de jeux de données.

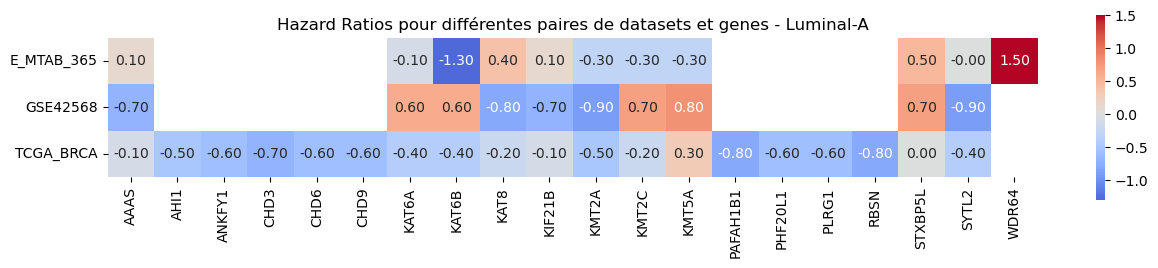
Nous avons ensuite construit le tableau suivant dont voici un extrait :



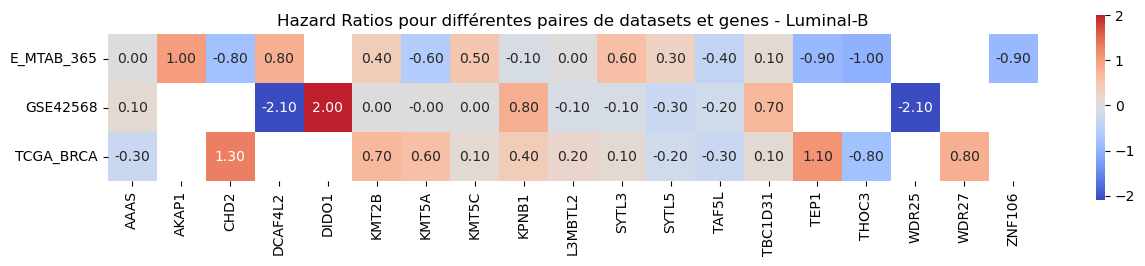
Pour visualiser les gènes que nous retrouvons dans le plus grand nombre de jeux de données, nous pouvons réaliser une matrice. Ce visuel propose les jeux de données en ligne et les gènes en colonnes. Les cellules de cette matrice sont colorées selon leur hasard ratio. En bleu, nous avons les croisements pour lesquels le hasard ratio est négatif et en rouge celle pour lesquels le hasard ratio est positif.

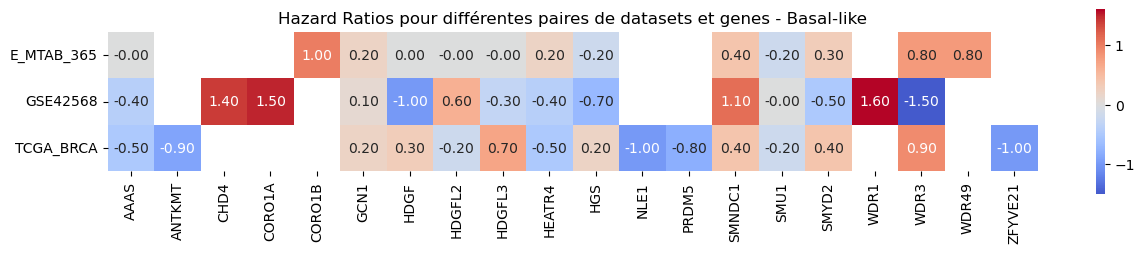
Voici la matrice que l’on obtient :

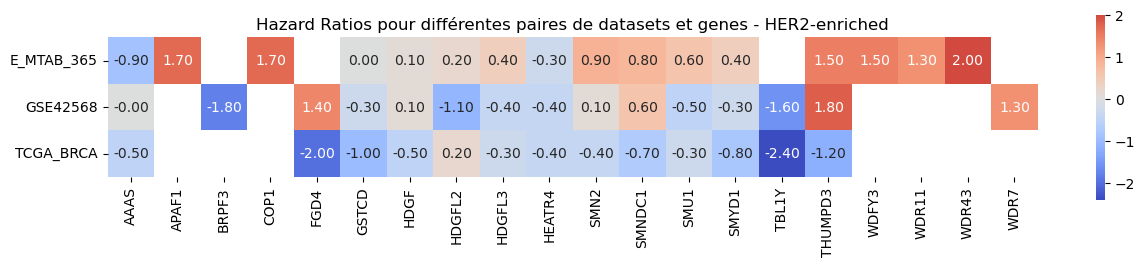
Dans l’exemple ci-dessus, nous avons représenté les gènes uniquement pour le sous type « All-tumours ». Nous pouvons représenter la même matrice pour tous les types.

Voici la matrice obtenue pour le sous type « Luminal-A » :

Dans la matrice ci-dessus, il y a beaucoup de données manquantes.

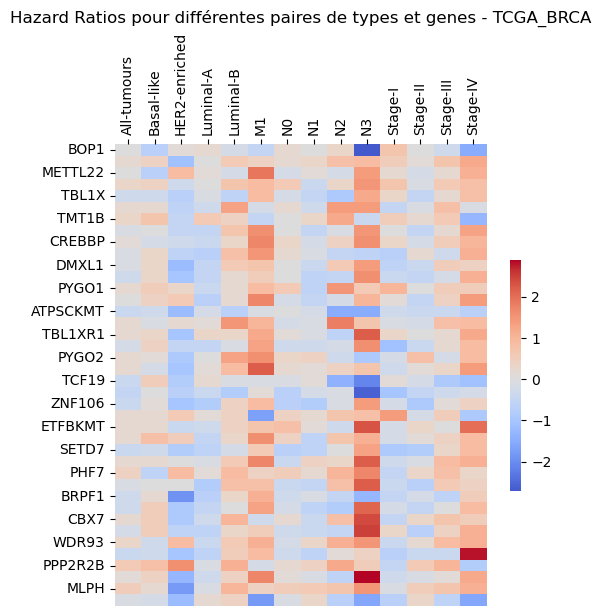
La matrice pour le sous type « Luminal-B » :

La matrice pour le sous type « Basal-like » :

Et enfin la matrice pour le sous type « HER2-enriched » :

**Bonus : Matrice des gènes VS les sous-types**

Pour le dataset TCGA-BRCA :



**Partie sur le potentiel métastasique**

Dans cette partie, nous allons essentiellement réaliser des volcano plot qui vont nous permettre d’identifier les gènes qui peuvent être à l’origine du développement de métastases.

