Le tableau général et les colonnes (1/2)

Le tableau entier contenant les 12151 gènes
NB : (12152 lignes car il y'a en plus aussi l'ordonnée à l'origine qui est calculée comme un feature appelé "*_SyntheticFeat4Intercept*")

- La colonne "Features" : les noms des gènes

La partie pour l'estimation de la présence ou absence du gène à travers les 4 sélections :

- Les colonnes ("*FS_Remagus04*", "*FS_MDAnderson_part1of2_310*", "*FS_BMS_Horak_2013*", "*FS_MT_3_dts*") sont de valeur 1 ou 0 pour dire le gène est présent dans la sélection ou pas; la colonne "*Percentage_selection_among_top4FSs*" donne le pourcentage de présence dans les 4 sélections.

La partie pour l'estimation de la stabilité des signes de coefs du gène :

- La colonne "*high_string_status*" donne si le gène a toujours le même signe sur toutes les répétitions et cela sur sur toutes les sélections où il est présent (Y=Yes, N=No)
- La colonne "*s_all*" est un score de stabilité des signes du gène (min 0, max 1, il s'agit du ratio de fois où le signe majoritaire est observé)
- La colonne "*s_all_abs*" est la valeur absolue de "s_all" et donc permet d'ordonner selon la stabilité l'ensemble des gènes, independamment du signe du gène
- La colonne "*SIG_status*" est de valeur 1 ou 0 pour dire si le gène est un SIG ou non. Un gène est un SIG si "*high_string_status*"=Y et "*s_all_abs*" = 1

Le tableau général et les colonnes (2/2)

La partie pour l'estimation des ranks des gènes :

- Les colonnes ("*Mean_coefs_rank_for_R04*", "*Mean_coefs_rank_for_MDA*" , "*Mean_coefs_rank_for_BMS*") sont pour chacune des sélections single-task le rank moyen obtenu sur les 10 répétitions. La colonne "*Mean_coefs_rank_for_all_3_STs*" est la moyenne de ces 3 colonnes.
- Les colonnes ("*Mean_coefs_rank_for_MT_copy_of_the_gene_from_R04*", "*Mean_coefs_rank_for_MT_copy_of_the_gene_from_MDA*" , "*Mean_coefs_rank_for_MT_copy_of_the_gene_from_BMS*") sont pour une "copie" du gène venant d'une cohorte et présent dans le multitask, le rank moyen obtenu sur les 10 répétitions. Le multitask utilise ici 3 cohortes donc 3 copies seront présentes si un gène est sélectionné dans le multitask. La colonne "*Mean_coefs_rank_for_MT_all_3_copies*" est la moyenne de ces 3 copies calculée en utilisant les 3 colonnes propres aux copies.

NB : Précédemment, nous avions une moyenne qui reunit les valeurs des 3 single-task et la moyenne sur les copies du multitask. Pour d'abord voir un état des choses avant d'avancer, je choisis de garder séparer ce qui vient du single-task et du multi-task. Ainsi donc, j'ai :

- **une moyenne pour les 3 colonnes single-task ("*Mean_coefs_rank_for_all_3_STs*")**
- **une moyenne pour les 3 copies du multitask ("*Mean_coefs_rank_for_MT_all_3_copies*")**

Ceci permet de ne pas mélanger pour le moment le rang sur le single-task qui est fait sur 12151 gènes et le rang sur le multi-task fait sur 36353 gènes (ie 12151 gènes par cohorte x 3 cohortes réunies).

Les différents tableaux enregistrés

- La sheet "*All_sortedby_STmean1*" est le tableau entier contenant les 12151 gènes, ordonné par la colonne ("*Mean_coefs_rank_for_all_3_STs*") ie du gène ayant le rang le plus petit au gène ayant le rang le plus élevé (l'idée est de classer selon ce qui ressort du single-task

- La sheet "*All_sortedby_MTmean1*" est le tableau entier contenant les 12151 gènes, ordonné par la colonne ("*Mean_coefs_rank_for_MT_all_3_copies*") ie du gène ayant le rang le plus petit au gène ayant le rang le plus élevé (l'idée est de classer selon ce qui ressort du multi-task

- Les sheets "*All_sortedby_STmean_2*" et "*All_sortedby_MTmean2*" sont les deux sheets précédentes, mais restreintes à seulement les colonnes les plus informatives globalement. Nous y avons :
* les colonnes de la partie pour l'estimation de la présence ou absence du gène à travers les 4 sélections
* les colonnes de la partie pour l'estimation de la stabilité des signes de coefs du gène
* 2 colonnes pour la moyenne des rangs (celle sortant du single-task, et celle sortant du multi-task)

Formatage pour lecture facilitée :
1 - En mettant la partie pour l'estimation de la stabilité des signes proches des ranks dans les sheets "...mean2" , on peut assez vite voir pourquoi certains des gènes avec un bon rang disparaissent en appliquant des restriction sur la stabilité.
2 - En fond rouge, sont les lignes où il y'a un gène avec une valeur dans les critères de bonne stabilité qui va le faire éliminer plus tard (ie lorsque "*high_string_status*" != Y et "*s_all_abs*" < 1)
3 - En fond vert est le titre des colonnes qui donnent le rang qui ressort globalement (sur le single-task ou le multitask). Les colonnes dont le titre est en orange sont les valeurs dont la moyenne a été faite.
4 - la colonne dont le titre est en gras est celle utilisée pour ordonner le tableau.

Un problème remarqué sur les rangs comptant tous les gènes...

- Un gène devrait avoir un rang assez proche sur des sélections différentes où il est présent.. Cependant, en classant le tableau de l'ensemble de nos gènes selon le rang dans une des 3 sélections single-task, on observe que le rang d'un gène dans les 2 autres sélections peut tout à fait être très variable.

Les 4 prochaines slides sont des captures d'écrans pour illustrer cela…

Cela veut dire qu'avec les valeurs de rangs sur tous les gènes, il est normal que nous ayons dans notre liste de gène finale obtenue en faisant la moyenne, des gènes en tête de liste qui ont un rang très éloigné de 1 celui-ci est issu de la moyenne de plusieurs rangs qui n' étaient pas très proches entre eux...

Une solution serait de revenir à des rangs utilisant une certaine normalisation…
Donnez-moi votre avis…
De mon coté, j'essaye de voir ce qui marche le mieux coté normalisation tout en montrant le classement dans l'ensemble des gènes…

- Nous classons ici selon le rang sur la sélection single-task de **Remagus04**
- et nous observons le haut de notre liste qui est probablement ce que l'on aimerait garder comme gènes intéressants
=> sur les 2 autres sélections single-task, les rangs peuvent être très différents...

Mean_coefs_rank_for_R04	Mean_coefs_rank_for_MDA	Mean_coefs_rank_for_BMS
4.2	1547.2	812.4
15.6	3272.8	3352.2
19.5	5436.1	14.0
19.6	2900.2	1579.8
20.2	8899.3	4678.0
21.6	995.8	20.4
26.4	2911.2	2898.8
29.4	1754.1	954.9
31.7	10005.0	9645.5
32.1	6296.3	4793.5
84.6	655.3	946.3

en nombres de gènes, taille FS / taille cohorte :	12151 / 12151	260 / 12151	12151 / 12151

colonne de rang utilisée pour ordonner le tableau du gène le plus important au gène le moins important

valeurs qui restent cohérentes à travers avec la sélection classée ici

valeurs qui sont incohérentes et donc vont avoir un impact sur le classement final avec une moyenne de ces rangs actuels

- Nous classons ici selon le rang sur la sélection single-task de **MDA**
- et nous observons le haut de notre liste qui est probablement ce que l'on aimerait garder comme gènes intéressants
=> la tête de liste de MDA est juste très différente de ce qui est sélectionné sur les deux autres cohortes...

Mean_coefs_rank_for_R04	Mean_coefs_rank_for_MDA	Mean_coefs_rank_for_BMS
8134.7	1.1	4692.0
4469.4	5.3	4555.7
4520.3	6.0	2924.7
9192.4	6.7	7556.3
6820.6	9.5	9278.9
4167.7	12.7	2800.6
3523.9	13.2	2877.9
11464.5	13.7	10042.3
2777.4	14.5	2698.5
7760.9	14.9	6218.7
4334.5	17.3	5903.1
2633.5	18.1	1919.1
4776.4	18.1	3762.1
9074.7	24.6	11564.4
2707.6	25.4	4388.0
929.7	29.3	813.7
5032.3	31.4	4153.0
3577.4	32.3	2543.6

en nombres de gènes, taille FS / taille cohorte :	12151 / 12151	260 / 12151	12151 / 12151

★ colonne de rang utilisée pour ordonner le tableau du gène le plus important au gène le moins important

valeurs qui restent cohérentes à travers avec la sélection classée ici

valeurs qui sont incohérentes et donc vont avoir un impact sur le classement final avec une moyenne de ces rangs actuels

- Nous classons ici selon le rang sur la sélection single-task de **BMS**
- et nous observons le haut de notre liste qui est probablement ce que l'on aimerait garder comme gènes intéressants
=> BMS et Remagus04 ont des similarités en tête de liste cependant des lots entiers de gènes viennent briser la cohérence des rangs en comparant les deux sélections...

Mean_coefs_rank_for_R04	Mean_coefs_rank_for_MDA	Mean_coefs_rank_for_BMS
1946.7	801.5	1.1
7501.3	11847.7	5.6
3449.4	4323.8	6.4
3860.7	2839.7	12.3
19.5	5436.1	14.0
345.4	2639.4	14.4
416.7	2668.2	15.8
21.6	995.8	20.4
1825.4	228.4	171.2
1328.8	369.4	232.5
2079.0	2251.5	236.2
2091.7	664.4	272.8
857.1	983.1	287.4
335.4	99.4	311.3
233.4	694.3	314.6
2058.7	103.4	344.3
en nombres de gènes, taille FS / taille cohorte :	12151 / 12151	260 / 12151	12151 / 12151

★ colonne de rang utilisée pour ordonner le tableau du gène le plus important au gène le moins important

valeurs qui restent cohérentes à travers avec la sélection classée ici

valeurs qui sont incohérentes et donc vont avoir un impact sur le classement final avec une moyenne de ces rangs actuels

- Nous classons ici **selon le rang moyen des 3 sélections single-task**
- et nous observons le haut de notre liste qui est probablement ce que l'on aimerait garder comme gènes intéressants

=> Nous avons ici 2 cas de valeurs incohérentes avec les autres sélections qui :
- l'une viendra augmenter le rang moyen
- et l'autre viendra diminuer le rang moyen

Mean_coefs_rank_for_R04	Mean_coefs_rank_for_MDA	Mean_coefs_rank_for_BMS	Mean_coefs_rank_for_all_3_STs
335.4	99.4	311.3	248.7000
21.6	995.8	20.4	345.9333
564.6	98.4	517.6	393.5333
233.4	694.3	314.6	414.1000
184.3	737.4	417.6	446.4333
866.0	284.4	344.9	498.4333
453.1	594.4	531.8	526.4333
138.6	639.5	882.7	553.6000
84.6	655.3	946.3	562.0667
750.3	232.4	719.8	567.5000
300.5	328.4	1085.7	571.5333
756.9	147.4	814.7	573.0000
509.9	616.4	603.3	576.5333
929.7	29.3	813.7	590.9000
962.7	187.4	717.5	622.5333
1337.4	144.4	426.4	636.0667

en nombres de gènes, taille FS / taille cohorte :	12151 / 12151	260 / 12151	12151 / 12151

colonne de rang utilisée pour ordonner le tableau du gène le plus important au gène le moins important

valeurs qui restent cohérentes à travers avec la sélection classée ici

valeurs qui sont incohérentes et donc vont avoir un impact sur le classement final avec une moyenne de ces rangs actuels