14/02/2022

tomorrow: meeting on the planning of the internship

I received pdfs from claire:

- Projet_GIRONDE_synopsis23052019.pdf
- 2021_Christinat_HRD_Oncoscan.pdf
- modpathol20153.pdf
- Nanocind_signature_S._CROCE.pdf

See notes_on_articles.md .
About the synopsis:

Projet_GIRONDE_synopsis23052019.pdf

le pronostic des lésions musculaires tumorales peut passer par des techniques comme l'index génomique (GI) = degré de complexité moléculaire & d'instabilité génomique. bon prédicteur de l'agressivité d'une tumeur. cela est fait à partir de la technologie Agilent, mais l'équipe veut étendre ça à la technologie Oncoscan, plus récente. pour cela, il est important de comparer ces deux technologies sur cette même technique. Cela est pertinent :

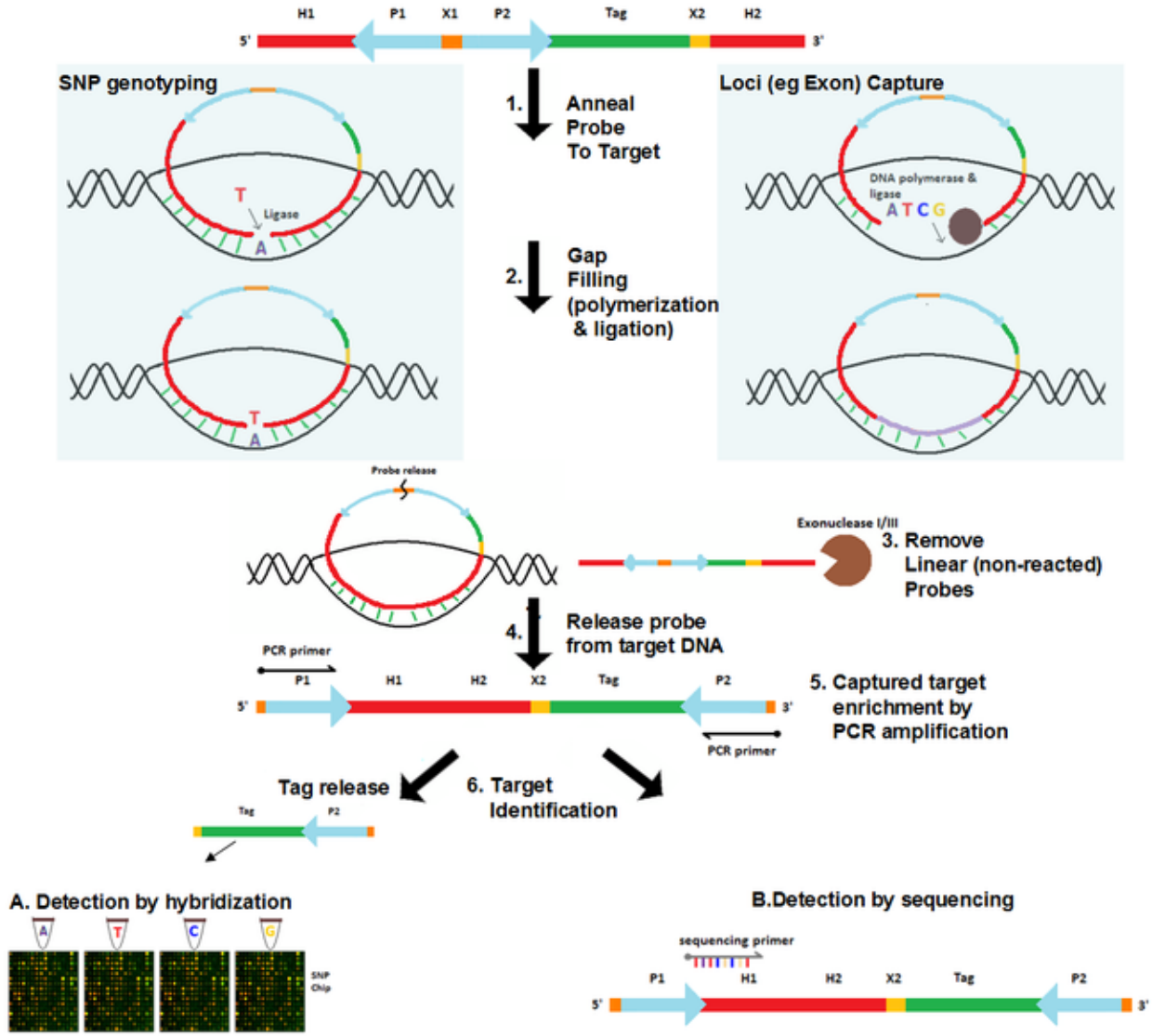
- en terme de diagnostique, car certaines tumeurs ne peuvent pas être traitées par les méthodes morphologiques.
- en terme de pronostique et de thérapeutique, cette méthode pourrait être appliquée à différentes tumeurs, même si on travaillera dans un premier temps sur les tumeurs stromales gastro-intestinales (GIST).

obj: transposer le calcul du GI et détermination du seuil de classification des tumeurs from Agilent to Affymetrix. Oncoscan a une résolution et une couverture génomique plus large. cela devrait permettre une plus grande précision dans le diagnostique.

Suivre la manip cette semaine

Remise en contexte: qu'est-ce que cette manip cherche à faire? -> produire un index génomique -> comment est calculé l'index génomique? GI = A^2/C où A = total number of alterations (segmental gains and losses) et C = the number of involved chromosomes. -> C est déterminé par l'expérience -> A : voir les résultats.

15:50 J'ai suivi la manip avec Laetitia. Il semble qu'on fasse du SNP genotyping, et pas du Loci Capture. Cette image illustre la manipulation dans son ensemble:



D'où viennent les échantillons liquides d'ADN? Laetitia reçoit des lames colorées Hématoxyline Éosine Safran (HES) + un FFPE: bloc de cellules tumorales extrait chirurgicalement. à l'aide de la lame (issue du bloc), elle sait où récupérer les cellules les plus tumorales du bloc. l'ADN est ensuite extrait de ces cellules sous forme liquide. cet ADN est appelé l'ADNg dans le protocole. Après avoir ajouté des sondes MIP, on le chauffe à 95°C pour passer de double à simple brin, puis on descend à 58°C pendant 2h pour laisser les sondes MIP s'associer aux brins d'ADN. Cette association est appelée Annealing. La structure en anneau ainsi formée est centrée sur un nucléotide du brin d'ADN, le seul qui n'est pas couvert par la sonde. l'étape suivante est le gap filling. on sépare le résultat de l'annealing en 2: un tube recevra du AT, l'autre du CG. les gaps seront complétés de manière complémentaire. ultimement, cela nous apprendra... qu'est-ce que ça nous apprend? Bref, une exonucléase est ensuite ajoutée pour dégrader les brins d'ADN libres (cela inclut les sondes non accrochées) et ne garder ainsi que les anneaux. ces derniers sont ensuite clivés, ce qui donne la structure suivante:

%%%%###########@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

Où:
%%%%% = Site 1 de clivage
= tag correspondant à la séquence d'ADN
@@@@@ = région homologue à la séquence d'ADN visée.

Deux PCR sont ensuite effectuées pour multiplier les sondes, puis une digestion Hae III est appliquée pour cliver la séquence ADN du reste de la sonde:

%%%%##########@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

L'hybridation sur puce affymetrix a ensuite lieu. Le tag effectue cette hybridation. Ce dernier est spécifique de la région d'ADNg. Ainsi, le nombre de copies de ce tag permet de connaître le nombre de copies de la région à laquelle il correspond. Le statut allélique de ces copies pourra également être déterminé en fonction de si ces copies sont présentes sur la puce AT ou la puce GC (car on utilise une puce par paire de bases.)

TODO:

Comment fonctionne Oncoscan CNV? Je veux savoir:

- sous quelle forme apparaissent les résultats
- quel est le but de la méthode?
- qu'est-ce que le gap filling nous apprend?

Questions:

- l'annealing laisse la place à un nucléotide. pourtant plus tard on vient boucher les trous avec des AT ou des CG, spécifiquement. quid?
- pourquoi ne faire du gap filling que sur AT et GC quand on pourrait le faire sur les 4 nucléotides? parce qu'ils sont complémentaires, mais ça ne répond que partiellement à la question. En fait si: un SNP consiste en un A/T qui se transform en C/G. peu importe lequel de la paire, c'est comme ça il semblerait.

aujourd'hui, j'ai fait 9h30-18h00 avec 30 minutes de pause le midi. Donc 8h de travail.

15/02/2022

Je continue de suivre la manip avec Laetitia. ce matin, on s'est arrêtés après avoir fini la PCR 1. Pour comprendre le fonctionnement global de la manip, voir https://en.wikipedia.org/wiki/DNA_microarray , la vidéo explique très bien le fonctionnement.

PCR: chercher "youtube PCR" donne une très bonne vidéo.

1. on chauffe fort pour séparer les 2 brins d'ADN
2. à une température plus basse, on laisse les primers s'hybrider sur les séquences d'intérêt
3. on chauffe un peu pour que les Taq polymérases se fixent aux primers. Ces dernières répliquent les brins d'ADN aux régions concernées. les étapes 2 et 3 sont répétées pour plusieurs cycles, ce qui double le nombre de copies de à chaque étape, résultant en une augmentation exponentielle.

Utiliser la partie précédente dans l'écriture du rapport de stage.

Je commence à bien connaître le protocole. Le but final est de déterminer l'index génomique et je cite `/home/waren/Desktop/stage_M2/sent_by_claire/sujet_stage_projet_GIRONDE__copie_from_administration.pdf` : "bien appréhender les critères utilisés [en utilisant les puces Agilent] pour la détection des variants afin de pouvoir proposer et développer une approche automatisée [...][à partir des] données Affymetrix". je cite également `Projet_GIRONDE_synopsis23052019.pdf`: "L'objectif de cette étude est de

transposer le calcul de l'index génomique et la détermination du seuil de classification des tumeurs de la technologie Agilent à la technologie Affymetrix/Oncoscan".

—> comment l'index génomique est-il déterminé pour Agilent? GI = A^2/C où A = total number of alterations (segmental gains and losses) et C = the number of involved chromosomes. -> C est déterminé par le protocole. on travaille sur l'humain donc 23 paires de chromosomes. -> A est calculé par l'expérience.

—> comment la détermination du seuil de classification des tumeurs est-elle faite pour Agilent? parle-t-on de la valeur de 10 pour le GI?

TODO: lire la prise de notes que j'ai fait avec claire et élodie. noter les infos ici, et les choses à faire. arrivée à 8h30 ----> départ à 17h. 30 min pause midi. Donc 8h de travail.

16/02/2022

J'ai suivi la fin de la manip avec Laetitia. J'ai vu comment les résultats sortaient d'affymétrix (.CEL, .ARR, .DAT), et le logiciel qui est utilisé pour les traiter. Cependant, je vais utiliser autre chose pour traiter ces fichiers, certainement des packages R.

TODO:

- lire la prise de notes que j'ai fait avec claire et élodie. noter les infos ici, et les choses à faire.
- voir les infos que contiennent les .CEL, .DAT et .ARR
- chercher les logiciels/packages qui lisent ces fichiers
- faire un bon récap de la manip et éclaircir les points obscurs:
    - PCR double brin
    - SNP?
- ~~organiser les dossiers comme élodie l'a indiqué~~
- bloquer le Jeudi 3 mars à 11h: https://u-bordeaux-fr.zoom.us/j/82532998606?pwd=a0Q3aWZ3ZjZMdC9udXcxem85clJPUT09
- noter les mardis de 12h à 14h: faire un point avec Élodie et Claire (et Sabrina!)
- noter qu'au début de chaque semaine, je dois aller voir la team CGH pour savoir qui analyse les résultats de la semaine et quand.

J'ai obtenu les codes pour me connecter à un ordi fixe de l'institut. la question actuelle est: ai-je un mail @bordeaux.unicancer.fr ?

Un repository git est créé et a été cloné sur le PC fixe de bergonié et sur le mien. la version la plus avancé est cassebriques.

Prochaine étape: récupérer des .CEL , .DAT , .ARR, et les explorer avec un logiciel/package R que j'aurai trouvé. sinon relire la todo list. arrivée à 9h10 ----> départ à 17h10. 30 min pause midi. Donc 7h30 de travail

17/02/2022

TODO:

- lire la prise de notes que j'ai fait avec claire et élodie. noter les infos ici, et les choses à faire.

- voir les infos que contiennent les .CEL, .DAT et .ARR
- chercher les logiciels/packages qui lisent ces fichiers --> le package R de l'article.
- faire un bon récap de la manip et éclaircir les points obscurs:
    - ~~PCR double brin~~
      les 2 brins sont formés séparément come le montre la vidéo, et se lient l'un à l'autre pour former le produit final double brin.
    - SNP?
- ~~organiser les dossiers comme élodie l'a indiqué~~
- bloquer le Jeudi 3 mars à 11h: https://u-bordeaux-fr.zoom.us/j/82532998606?pwd=a0Q3aWZ3ZjZMdC9udXcxem85clJPUT09
- noter les mardis de 12h à 14h: faire un point avec Élodie et Claire (et Sabrina!)
- noter qu'au début de chaque semaine, je dois aller voir la team CGH pour savoir qui analyse les résultats de la semaine et quand.
- ~~lire en profondeur l'article arm-level.~~ CCL & discussion: "notre méthode est aussi efficace que les experts humains et bien plus rapide."
    - savoir expliquer cette méthode et comment l'utiliser.
    - prendre des données auprès de l'équipe technique. apporter une clé USB demain! récupérer les .ARR, .DAT et surtout .CEL .
    - l'utiliser sur nos données.

Ai installé R et Rstudio sur Bergonié, ai téléchargé le package R de l'article arm-level à https://github.com/yannchristinat/oncoscanR-public. faire un push propre sur cass et (tenter de) le pull sur bergonié. le push c'est bon mais bergonié ne peut même pas faire de commits. la commande `git add` à elle seule trigger ce message:

```
C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie>git add
warning: unable to access 'P://.gitconfig': Permission denied
warning: unable to access 'P://.gitconfig': Permission denied
warning: unable to access 'P://.gitconfig': Permission denied
fatal: unknown error occurred while reading the configuration files
```

voir la réponse de jennifer pour ça. potentiellement demander un linux

arrivée à 10h -> départ à 16h50 = 6h20 de travail total cumulé sur la semaine: 29h50. pour faire 35h, reste 5h10.

18/02/2022

ai résolu le pb qui m'empêchait de pull en supprimant la variable d'envt HOMEPATH: https://stackoverflow.com/questions/14774159/git-warning-unable-to-access-p-gitconfig-invalid-argument

Ai fait un push clean. je travaille maintenant essentiellement sur bergo.

TODO:

- lire la prise de notes que j'ai fait avec claire et élodie. noter les infos ici, et les choses à faire.
- faire un bon récap de la manip et éclaircir les points obscurs:

    - ~~PCR double brin~~
      les 2 brins sont formés séparément come le montre la vidéo, et se lient l'un à l'autre pour former le produit final double brin.
    - SNP?
- ~~organiser les dossiers comme élodie l'a indiqué~~
- bloquer le Jeudi 3 mars à 11h: https://u-bordeaux-fr.zoom.us/j/82532998606?pwd=a0Q3aWZ3ZjZMdC9udXcxem85clJPUT09
- noter les mardis de 12h à 14h: faire un point avec Élodie et Claire (et Sabrina!)
- noter qu'au début de chaque semaine, je dois aller voir la team CGH pour savoir qui analyse les résultats de la semaine et quand.
- ~~voir les infos que contiennent les .CEL, .DAT et .ARR~~
- ~~chercher les logiciels/packages qui lisent ces fichiers~~ --> le package R de l'article.
- ~~lire en profondeur l'article arm-level.~~ CCL & discussion: "notre méthode est aussi efficace que les experts humains et bien plus rapide."
    - savoir expliquer cette méthode et comment l'utiliser.
    - ~~prendre des données auprès de l'équipe technique. apporter une clé USB demain! récupérer les .ARR, .DAT et surtout .CEL .~~
    - l'utiliser sur nos données.

.ARR = du xml

Laetitia m'a passé les données anonymisées d'affymetrix. Je fais passer le premier échantillon dans le package R.

path to R.exe: "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\R.exe" path to Oncoscan-R script: "C:\Users\e.bordron\Documents\R\R-4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R" path to first txt file: "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\1-RV.OSCHP.segments.txt"

to set an environment variable:

```
setx r_exe "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\R.exe"
```

to use it:

```
%r_exe%
```

output:

```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.
```

```
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.
>
```

commands entered to create envt variables:

> setx r_exe "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\Rscript.exe"
> setx oncos-r "C:\Users\e.bordron\Documents\R\R-4.1.2\library\oncoscanR\bin\oncoscan-workflow.R"
> setx data_folder "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data"\

command to run workflow:

> r_exe

Unrelated but the theme Shades of Purple is cool but for markdown editing I prefer built-in theme Monokai Dimmed. Also on Bergo I save my keybindings.json file in C:\Users\e.bordron\Documents .

I created 2 folders for data: raw_data, a backup folder, and working_data, which I will be working on. Before doing that, I unintentionaly lost the file segments.txt for the first sample by sending the output of a command into it.

I did this:

```
> "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\Rscript.exe"
"C:\Users\e.bordron\Documents\R\R-
4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R"
"C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\1-
RV.OSCHP.segments.txt" M

Erreur dans UseMethod("collector_value") :
  pas de méthode pour 'collector_value' applicable pour un objet de classe
"c('collector_skip', 'collector')"
Appels : workflow_oncoscan.run ... load_chas -> read_tsv -> <Anonymous> ->
collector_value
De plus : Message d'avis :
The following named parsers don't match the column names: CN State, Type, Full
Location
Exécution arrêtée
```

Prochaine chose à faire: modifier les colonnes du .txt pour qu'il y ait seulement le s3 dont le package a besoin.

nouveau problème quand je lance cette ligne de commande:

> "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\Rscript.exe" "C:\Users\e.bordron\Documents\R\R-4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R" "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\woring_data\2-AD\2-ADREC.RC.OSCHP.segments.txt" F

> Accès refusé.

et une pop-up apparaît:

```
Cette application ne peut pas s'exécuter sur votre PC.
```

même quand je vais dans C:\Users\e.bordron\Documents\R\R-4.1.2\bin et que je fais:

> Rscript.exe

le même problème survient. essayer de redémarrer.

Je viens de redémarrer. je me rends compte sur `https://helpdeskgeek.com/windows-10/how-to-fix-this-app-cant-run-on-your-pc-in-windows-10/` que Rscript.exe est en 32 bit. edit: j'ai les 2 Rscript: le 32bit et le 64bit. le chemin du 32bit: `C:\Users\e.bordron\Documents\R\R-4.1.2\bin` le chemin du 64bit: `C:\Users\e.bordron\Documents\R\R-4.1.2\bin\x64`

J'ai le problème suivant, en tapant la bonne ligne de commande:

> "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\x64\Rscript.exe" "C:\Users\e.bordron\Documents\R\R-4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R" "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\working_data\2-AD\2-ADREC.RC.OSCHP.segments.txt" F

```
Erreur dans load_chas(chas.fn, oncoscan.cov) : Parsing ChAS file failed.
Appels : workflow_oncoscan.run -> load_chas
De plus : Message d'avis :
The following named parsers don't match the column names: Full Location
Exécution arrêtée
```

on en revient à cette solution: Prochaine chose à faire: modifier les colonnes du .txt pour qu'il y ait seulement les 3 dont le package a besoin: `Type, CN State and Full Location` Je viens d'essayer: les colonnes Type et CN state sont bien présentes, mais pas la colonne Full Location. J'aurais du lui dire plus tôt, mais je vais demander à Laetita si il est possible d'avoir cette 3eme colonne. Si ce n'est pas possible, peut-être qu'il est possible de récupérer cette information à partir d'autres colonnes, auquel cas je regarderai d'autres moyens d'obtenir des données de ces .CEL. parser moi-même ce fichier est aussi une option

arrivée à 10h10 -> 17:30 = 6h50 de travail et 30 min de pause le midi. 1h40 est déjà faite pour la semaine prochaine.

21/02/2022

Je regarde vite fait si l'une des colonnes contient par hasard l'information "Full location", sinon je demande à Laetitia. La colonne `Full location` semble être un arrondi au 100 des valeurs de position contenues dans `Microarray Nomenclature`. Je fais une colonne Full location à partir de ça. en R:

- import CSV as dataframe
- get value between parenthesis: `(754,192-145,095,477)` from column `Microarray Nomenclature`
- get chromosome from column `Chromosome`

- create column `Full location` that contains such values: `chr7:129199300-129813700`

fait. je lance oncoscan-R dessus:

> "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\x64\Rscript.exe" "C:\Users\e.bordron\Documents\R\R-4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R" "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\working_data\2-AD\2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt" F

Cela me donne une erreur:

```
Erreur dans load_chas(chas.fn, oncoscan.cov) : Parsing ChAS file failed.
Appels : workflow_oncoscan.run -> load_chas
De plus : Message d'avis :
The following named parsers don't match the column names: Full Location
Exécution arrêtée
```

je le lance aussi dans R. voir csv_formatting.R. le message d'erreur est:

```
Error in load_chas(chas.fn, oncoscan.cov) : Parsing ChAS file failed.
In addition: Warning message:
The following named parsers don't match the column names: Full Location
```

Or, la colonne Full location est bien écrite.

résolu. maintenant:

> workflow_oncoscan.run("C:/Users/e.bordron/Desktop/CGH-scoring/M2_internship_Bergonie/data/working_data/2-AD/2-ADREC.RC.OSCHP.segments_FULL_LOCA ..." ... [TRUNCATED]

```
Error in if (length(parm) == 0 || seg_start > end(parm)) {:

    missing value where TRUE/FALSE needed
In addition: Warning messages:
1: In load_chas(chas.fn, oncoscan.cov) : NAs introduced by coercion
2: In load_chas(chas.fn, oncoscan.cov) : NAs introduced by coercion
```

arrivée à 10h25, départ à 19:25

22/02/2022

obj du jour: lancer l'analyse par oncoscanR et obtenir des résultats.

```
Error in seg_cntype %in% c(cntype.gain, cntype.loss, cntype.loh) :
  object 'cntype.gain' not found
```

--> pour éviter cette erreur, faire library(oncoscanR), même si la commande est appelée avec la syntaxe "oncoscanR::workflow_oncoscan.run(path, gender)".

l'analyse fonctionne. en script:

> oncoscanR::workflow_oncoscan.run("C:/Users/e.bordron/Desktop/CGH-scoring/M2_internship_Bergonie/data/working_data/2-AD/2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt", "F")

```
$armlevel
$armlevel$AMP
character(0)

$armlevel$LOSS
[1] "14q" "15q" "17p" "1p"  "22q" "3q"

$armlevel$LOH
[1] "17q" "18q" "21q"

$armlevel$GAIN
[1] "5p" "5q"


$scores
$scores$LST
[1] 0

$scores$LOH
[1] 5

$scores$TDplus
[1] 0


$gender
[1] "F"

$file
[1] "2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt"
```

en ligne de commande:

> "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\x64\Rscript.exe" "C:\Users\e.bordron\Documents\R\R-4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R" "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\working_data\2-AD\2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt" F

```
{
  "armlevel": {
    "AMP": [],
    "LOSS": ["14q", "15q", "17p", "1p", "22q", "3q"],
    "LOH": ["17q", "18q", "21q"],
    "GAIN": ["5p", "5q"]
  },
  "scores": {
    "LST": 0,
    "LOH": 5,
    "TDplus": 0
  },
  "gender": "F",
  "file": "2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt"
}
```


I set environment variables on bergo:

```
setx oncos-r "C:\Users\e.bordron\Documents\R\R-
4.1.2\library\oncoscanR\bin\run_oncoscan_workflow.R"
setx r_exe "C:\Users\e.bordron\Documents\R\R-4.1.2\bin\x64\Rscript.exe"
```

la ligne de commande plus courte est donc:

> %r_exe% %oncos-r% "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\working_data\2-AD\2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt" F.

au fait, faire `set` permet d'afficher les variables d'environnement. et créer une variable avec setx nécessite de rouvrir le terminal pour pouvoir l'utiliser.

Comment analyser ces données? l'index génomique est calculé à partir du (nombre d'altérations)^2 / nombre de chromosomes affectés.

Aussi tester EaCoN (Easy Copy Number).

!! Pour info, Dans windows -> Options internet -> avancé, j'ai décoché la case "utiliser TLS 1.2" sur bergo. !!

I am currently trying to install EaCoN on R. I follow the install instructions at https://github.com/gustaveroussy/EaCoN. I installed the `Core` part, including BiocManager from R's menus. In `MICROARRAY-SPECIFIC` part, I install the `ONCOSCAN FAMILY (OncoScan / OncoScan_CNV)` part:

> devtools::install_github("gustaveroussy/apt.oncoscan.2.4.0")

```
Downloading GitHub repo gustaveroussy/apt.oncoscan.2.4.0@HEAD
Error in utils::download.file(url, path, method = method, quiet = quiet,  :
  cannot open URL
'https://api.github.com/repos/gustaveroussy/apt.oncoscan.2.4.0/tarball/HEAD'
```

I do (removing the s from https):

> utils::download.file('http://api.github.com/repos/gustaveroussy/apt.oncoscan.2.4.0/tarball/HEAD', "C:/Users/e.bordron/Desktop/CGH-scoring/M2_internship_Bergonie/results/delme")

it doesn't work. 😕

edit: after going to Windows -> Options internet -> avancé -> tout en bas: cocher la case "Utiliser TLS 1.2", ça a marché:

> devtools::install_github("gustaveroussy/apt.oncoscan.2.4.0")

```
Downloading GitHub repo gustaveroussy/apt.oncoscan.2.4.0@HEAD
Skipping 1 packages not available: affxparser
√  checking for file
'C:\Users\e.bordron\AppData\Local\Temp\RtmpmOFoai\remotes7083e8f37f4\gustaveroussy
-apt.oncoscan.2.4.0-e14fca3/DESCRIPTION' (426ms)
-  preparing 'apt.oncoscan.2.4.0': (649ms)
√  checking DESCRIPTION meta-information ...
-  checking for LF line-endings in source and make files and shell scripts
-  checking for empty or unneeded directories
   Omitted 'LazyData' from DESCRIPTION
-  building 'apt.oncoscan.2.4.0_0.1.6.tar.gz'

* installing *source* package 'apt.oncoscan.2.4.0' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
  converting help for package 'apt.oncoscan.2.4.0'
    finding HTML links ... done
    apt_oncoscan_process                    html
    apt_oncoscan_process_batch              html
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (apt.oncoscan.2.4.0)
```

next step: continue installing from Eacon github.

arrivée à 10h30, pas de pause le midi. départ à 15h30

23/02/2022

J'installe l'annotation pour le `NA33 (hg19) build` pour le `OncoScan_CNV design` :

> install.packages("https://zenodo.org/record/5494853/files/OncoScanCNV.na33.r2_0.1.0.tar.gz", repos = NULL, type = "source")

```
trying URL
'https://zenodo.org/record/5494853/files/OncoScanCNV.na33.r2_0.1.0.tar.gz'
Content type 'application/octet-stream' length 170586701 bytes (162.7 MB)
downloaded 162.7 MB

* installing *source* package 'OncoScanCNV.na33.r2' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
No man pages found in package  'OncoScanCNV.na33.r2'
*** installing help indices
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (OncoScanCNV.na33.r2)
```

Tout va bien.

et l'annotation pour le `NA36 (hg38) build` pour le `OncoScan_CNV design` :

> install.packages("https://zenodo.org/record/5494853/files/OncoScanCNV.na36.r1_0.1.0.tar.gz", repos = NULL, type = "source")

```
trying URL
'https://zenodo.org/record/5494853/files/OncoScanCNV.na36.r1_0.1.0.tar.gz'
Content type 'application/octet-stream' length 195992364 bytes (186.9 MB)
downloaded 23.9 MB

Warning in install.packages :
  downloaded length 25047040 != reported length 195992364
Erreur dans untar2(tarfile, files, list, exdir, restore_times) :
  incomplete block on file
Warning in install.packages :
  installation of package
```

```
'C:/Users/EC817~1.BOR/AppData/Local/Temp/RtmpgHYxuC/downloaded_packages/OncoScanCN
V.na36.r1_0.1.0.tar.gz' had non-zero exit status
```

c'est un problème. Je réessaie:

```
trying URL
'https://zenodo.org/record/5494853/files/OncoScanCNV.na36.r1_0.1.0.tar.gz'
Content type 'application/octet-stream' length 195992364 bytes (186.9 MB)
downloaded 186.9 MB

* installing *source* package 'OncoScanCNV.na36.r1' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
No man pages found in package  'OncoScanCNV.na36.r1'
*** installing help indices
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (OncoScanCNV.na36.r1)
```

tout a l'air d'aller. j'essaie la partie `USAGE`, `Step by step mode`, `Raw data processing`, `Affymetrix OncoScan / OncoScan_CNV`: voir eacon.R

I did this:

```
library(EaCoN)
pathToATCelFile = "C:/Users/e.bordron/Desktop/CGH-
scoring/M2_internship_Bergonie/data/working_data/2-AD/2-AD_AT_(OncoScan_CNV).CEL"
pathToGCCelFile = "C:/Users/e.bordron/Desktop/CGH-
scoring/M2_internship_Bergonie/data/working_data/2-AD/2-AD_GC_(OncoScan_CNV).CEL"
outputFolder = "2-AD"
OS.Process(ATChannelCel = pathToATCelFile, GCChannelCel = pathToGCCelFile,
samplename = "2-AD")
```

output:

```
'getOption("repos")' replaces Bioconductor standard repositories, see '?
repositories' for details
```

```
replacement repositories:
    CRAN: https://cran.rstudio.com/

 [PC2979:10580] BSgenome BSgenome.Hsapiens.UCSC.hg19 available but not installed.
Please install it !
Error:
```

Je dois installer ce génome à l'aide du github, partie `INSTALLATION`, `GENOMES`.

```
BiocManager::install('BSgenome.Hsapiens.UCSC.hg19')
```

```
'getOption("repos")' replaces Bioconductor standard repositories, see '?
repositories' for details

replacement repositories:
    CRAN: https://cran.rstudio.com/

Bioconductor version 3.14 (BiocManager 1.30.16), R 4.1.2 (2021-11-01)
Installing package(s) 'BSgenome.Hsapiens.UCSC.hg19'
installation du package source ‘BSgenome.Hsapiens.UCSC.hg19’

trying URL
'https://bioconductor.org/packages/3.14/data/annotation/src/contrib/BSgenome.Hsapi
ens.UCSC.hg19_1.4.3.tar.gz'
Content type 'application/x-gzip' length 710245413 bytes (677.3 MB)
downloaded 677.3 MB

* installing *source* package 'BSgenome.Hsapiens.UCSC.hg19' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
  converting help for package 'BSgenome.Hsapiens.UCSC.hg19'
    finding HTML links ... done
    package                                 html
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (BSgenome.Hsapiens.UCSC.hg19)

The downloaded source packages are in
    ‘C:\Users\e.bordron\AppData\Local\Temp\RtmpgHYxuC\downloaded_packages’
Old packages: 'cli'
Update all/some/none? [a/s/n]:
a
```

```
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/cli_3.2.0.zip'
Content type 'application/zip' length 1257150 bytes (1.2 MB)
downloaded 1.2 MB

package ‘cli’ successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package ‘cli’
Warning: restored ‘cli’

The downloaded binary packages are in
    C:\Users\e.bordron\AppData\Local\Temp\RtmpgHYxuC\downloaded_packages
Warning message:
In file.copy(savedcopy, lib, recursive = TRUE) :
  problem copying C:\Users\e.bordron\Documents\R\R-
4.1.2\library\00LOCK\cli\libs\x64\cli.dll to C:\Users\e.bordron\Documents\R\R-
4.1.2\library\cli\libs\x64\cli.dll: Permission denied
```

an error pop-up is triggered by this command. on `https://github.com/gustaveroussy/EaCoN/issues/16`, it is indicated that it is a known issue on windows, but it works on linux. So I guess it can't be used on windows in our case.

at the last meeting, this was said:

Objectifs

- ~Tester VMware workstation sur le PC de Bergonié et évaluer si les données peuvent-être traitées en local~ -> besoin droits admin, eacon ne marche pas avec linux a priori
- Lire et comprendre la documentation (manuel et papier) des packages oncoScanR, rCGH et EaCoN. Produire un tableau comparatif
- Quelles sont les fichiers d'entrées ?
- Quelles fonctionnalités sont disponibles ?
- Quels type d'output ? Plots, résultats (log-ratio, prédiction des CNV *sous quelle forme?* etc)
- Ecrire une fonction d'implémentation du GI

I tried to install VMware workstation Player and VMware workstation Pro. When I use them, both the installers warn me: `Vous avez besoin des privilèges administrateur pour installer ce logiciel.` This is a pop-up, I have no way to interact with them.

Je commence un tableau récap des 3 packages R (oncoScanR, rCGH et EaCoN). voir https://annuel2.framapad.org/p/CR_r%C3%A9unions_CGH_-_Elie pour plus d'infos.

OncoscanR : Computation of arm-level alteration. Method can be tweaked Score LST -> see Popova et al, Can. Res. 2012 (PMID: 22933060) -> Ploidy and Large-Scale Genomic Instability Consistently Identify Basal-like Breast Carcinomas with BRCA1/2 Inactivation Score LOH -> see Abkevich et al., Br J Cancer 2012 (PMID: 23047548) Score Tdplus -> see Popova et al., Cancer Res 2016 (PMID: 26787835)

j'installe rCGH. l'output est long mais je le mets ici:

```
BiocManager::install("rCGH")
```

```
> BiocManager::install("rCGH")
'getOption("repos")' replaces Bioconductor standard repositories, see '?
repositories' for details

replacement repositories:
    CRAN: https://cran.rstudio.com/

Bioconductor version 3.14 (BiocManager 1.30.16), R 4.1.2 (2021-11-01)
Installing package(s) 'rCGH'
installation des dépendances ‘assertthat’, ‘colorspace’, ‘dbplyr’, ‘filelock’,
‘farver’, ‘labeling’, ‘munsell’, ‘viridisLite’, ‘sass’, ‘blob’, ‘plogr’,
‘BiocFileCache’, ‘png’, ‘gtable’, ‘isoband’, ‘scales’, ‘httpuv’, ‘xtable’,
‘fontawesome’, ‘sourcetools’, ‘bslib’, ‘affyio’, ‘preprocessCore’, ‘DBI’,
‘RSQLite’, ‘biomaRt’, ‘KEGGREST’, ‘multtest’, ‘plyr’, ‘DNAcopy’, ‘ggplot2’,
‘shiny’, ‘affy’, ‘TxDb.Hsapiens.UCSC.hg18.knownGene’,
‘TxDb.Hsapiens.UCSC.hg19.knownGene’, ‘TxDb.Hsapiens.UCSC.hg38.knownGene’,
‘org.Hs.eg.db’, ‘GenomicFeatures’, ‘AnnotationDbi’, ‘aCGH’

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/assertthat_0.2.1.zip'
Content type 'application/zip' length 55017 bytes (53 KB)
downloaded 53 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/colorspace_2.0-3.zip'
Content type 'application/zip' length 2651585 bytes (2.5 MB)
downloaded 2.5 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/dbplyr_2.1.1.zip'
Content type 'application/zip' length 835602 bytes (816 KB)
downloaded 816 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/filelock_1.0.2.zip'
Content type 'application/zip' length 39993 bytes (39 KB)
downloaded 39 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/farver_2.1.0.zip'
Content type 'application/zip' length 1752630 bytes (1.7 MB)
downloaded 1.7 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/labeling_0.4.2.zip'
Content type 'application/zip' length 62679 bytes (61 KB)
downloaded 61 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/munsell_0.5.0.zip'
Content type 'application/zip' length 245248 bytes (239 KB)
downloaded 239 KB

trying URL
'https://cran.rstudio.com/bin/windows/contrib/4.1/viridisLite_0.4.0.zip'
Content type 'application/zip' length 1299509 bytes (1.2 MB)
downloaded 1.2 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/sass_0.4.0.zip'
Content type 'application/zip' length 3639310 bytes (3.5 MB)
```

```
downloaded 3.5 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/blob_1.2.2.zip'
Content type 'application/zip' length 48066 bytes (46 KB)
downloaded 46 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/plogr_0.2.0.zip'
Content type 'application/zip' length 18940 bytes (18 KB)
downloaded 18 KB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/BiocFileCache
_2.2.1.zip'
Content type 'application/zip' length 562298 bytes (549 KB)
downloaded 549 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/png_0.1-7.zip'
Content type 'application/zip' length 336780 bytes (328 KB)
downloaded 328 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/gtable_0.3.0.zip'
Content type 'application/zip' length 434251 bytes (424 KB)
downloaded 424 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/isoband_0.2.5.zip'
Content type 'application/zip' length 2726831 bytes (2.6 MB)
downloaded 2.6 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/scales_1.1.1.zip'
Content type 'application/zip' length 558382 bytes (545 KB)
downloaded 545 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/httpuv_1.6.5.zip'
Content type 'application/zip' length 1695904 bytes (1.6 MB)
downloaded 1.6 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/xtable_1.8-4.zip'
Content type 'application/zip' length 706535 bytes (689 KB)
downloaded 689 KB

trying URL
'https://cran.rstudio.com/bin/windows/contrib/4.1/fontawesome_0.2.2.zip'
Content type 'application/zip' length 1529199 bytes (1.5 MB)
downloaded 1.5 MB

trying URL
'https://cran.rstudio.com/bin/windows/contrib/4.1/sourcetools_0.1.7.zip'
Content type 'application/zip' length 691390 bytes (675 KB)
downloaded 675 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/bslib_0.3.1.zip'
Content type 'application/zip' length 5038570 bytes (4.8 MB)
downloaded 4.8 MB
```

```
trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/affyio_1.64.0
.zip'
Content type 'application/zip' length 170834 bytes (166 KB)
downloaded 166 KB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/preprocessCor
e_1.56.0.zip'
Content type 'application/zip' length 266353 bytes (260 KB)
downloaded 260 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/DBI_1.1.2.zip'
Content type 'application/zip' length 741964 bytes (724 KB)
downloaded 724 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/RSQLite_2.2.10.zip'
Content type 'application/zip' length 2545233 bytes (2.4 MB)
downloaded 2.4 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/biomaRt_2.50.
3.zip'
Content type 'application/zip' length 979894 bytes (956 KB)
downloaded 956 KB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/KEGGREST_1.34
.0.zip'
Content type 'application/zip' length 192380 bytes (187 KB)
downloaded 187 KB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/multtest_2.50
.0.zip'
Content type 'application/zip' length 979735 bytes (956 KB)
downloaded 956 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/plyr_1.8.6.zip'
Content type 'application/zip' length 1499372 bytes (1.4 MB)
downloaded 1.4 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/DNAcopy_1.68.
0.zip'
Content type 'application/zip' length 792263 bytes (773 KB)
downloaded 773 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/ggplot2_3.3.5.zip'
Content type 'application/zip' length 4130564 bytes (3.9 MB)
downloaded 3.9 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/shiny_1.7.1.zip'
Content type 'application/zip' length 4230764 bytes (4.0 MB)
```

```
downloaded 4.0 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/affy_1.72.0.z
ip'
Content type 'application/zip' length 1768641 bytes (1.7 MB)
downloaded 1.7 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/GenomicFeatur
es_1.46.4.zip'
Content type 'application/zip' length 2430857 bytes (2.3 MB)
downloaded 2.3 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/AnnotationDbi
_1.56.2.zip'
Content type 'application/zip' length 5387247 bytes (5.1 MB)
downloaded 5.1 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/aCGH_1.72.0.z
ip'
Content type 'application/zip' length 2728819 bytes (2.6 MB)
downloaded 2.6 MB

trying URL
'https://bioconductor.org/packages/3.14/bioc/bin/windows/contrib/4.1/rCGH_1.24.0.z
ip'
Content type 'application/zip' length 5180013 bytes (4.9 MB)
downloaded 4.9 MB

package ‘assertthat’ successfully unpacked and MD5 sums checked
package ‘colorspace’ successfully unpacked and MD5 sums checked
package ‘dbplyr’ successfully unpacked and MD5 sums checked
package ‘filelock’ successfully unpacked and MD5 sums checked
package ‘farver’ successfully unpacked and MD5 sums checked
package ‘labeling’ successfully unpacked and MD5 sums checked
package ‘munsell’ successfully unpacked and MD5 sums checked
package ‘viridisLite’ successfully unpacked and MD5 sums checked
package ‘sass’ successfully unpacked and MD5 sums checked
package ‘blob’ successfully unpacked and MD5 sums checked
package ‘plogr’ successfully unpacked and MD5 sums checked
package ‘BiocFileCache’ successfully unpacked and MD5 sums checked
package ‘png’ successfully unpacked and MD5 sums checked
package ‘gtable’ successfully unpacked and MD5 sums checked
package ‘isoband’ successfully unpacked and MD5 sums checked
package ‘scales’ successfully unpacked and MD5 sums checked
package ‘httpuv’ successfully unpacked and MD5 sums checked
package ‘xtable’ successfully unpacked and MD5 sums checked
package ‘fontawesome’ successfully unpacked and MD5 sums checked
package ‘sourcetools’ successfully unpacked and MD5 sums checked
package ‘bslib’ successfully unpacked and MD5 sums checked
package ‘affyio’ successfully unpacked and MD5 sums checked
```

```
package ‘preprocessCore’ successfully unpacked and MD5 sums checked
package ‘DBI’ successfully unpacked and MD5 sums checked
package ‘RSQLite’ successfully unpacked and MD5 sums checked
package ‘biomaRt’ successfully unpacked and MD5 sums checked
package ‘KEGGREST’ successfully unpacked and MD5 sums checked
package ‘multtest’ successfully unpacked and MD5 sums checked
package ‘plyr’ successfully unpacked and MD5 sums checked
package ‘DNAcopy’ successfully unpacked and MD5 sums checked
package ‘ggplot2’ successfully unpacked and MD5 sums checked
package ‘shiny’ successfully unpacked and MD5 sums checked
package ‘affy’ successfully unpacked and MD5 sums checked
package ‘GenomicFeatures’ successfully unpacked and MD5 sums checked
package ‘AnnotationDbi’ successfully unpacked and MD5 sums checked
package ‘aCGH’ successfully unpacked and MD5 sums checked
package ‘rCGH’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\e.bordron\AppData\Local\Temp\RtmpoXryBf\downloaded_packages
installation des packages sources ‘TxDb.Hsapiens.UCSC.hg18.knownGene’,
‘TxDb.Hsapiens.UCSC.hg19.knownGene’, ‘TxDb.Hsapiens.UCSC.hg38.knownGene’,
‘org.Hs.eg.db’

trying URL
'https://bioconductor.org/packages/3.14/data/annotation/src/contrib/TxDb.Hsapiens.
UCSC.hg18.knownGene_3.2.2.tar.gz'
Content type 'application/x-gzip' length 16032792 bytes (15.3 MB)
downloaded 15.3 MB

trying URL
'https://bioconductor.org/packages/3.14/data/annotation/src/contrib/TxDb.Hsapiens.
UCSC.hg19.knownGene_3.2.2.tar.gz'
Content type 'application/x-gzip' length 18669702 bytes (17.8 MB)
downloaded 17.8 MB

trying URL
'https://bioconductor.org/packages/3.14/data/annotation/src/contrib/TxDb.Hsapiens.
UCSC.hg38.knownGene_3.14.0.tar.gz'
Content type 'application/x-gzip' length 43518321 bytes (41.5 MB)
downloaded 41.5 MB

trying URL
'https://bioconductor.org/packages/3.14/data/annotation/src/contrib/org.Hs.eg.db_3
.14.0.tar.gz'
Content type 'application/x-gzip' length 82195112 bytes (78.4 MB)
downloaded 78.4 MB

* installing *source* package 'TxDb.Hsapiens.UCSC.hg18.knownGene' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
  converting help for package 'TxDb.Hsapiens.UCSC.hg18.knownGene'
```

```
    finding HTML links ... done
    package                                 html
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (TxDb.Hsapiens.UCSC.hg18.knownGene)
* installing *source* package 'TxDb.Hsapiens.UCSC.hg19.knownGene' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
  converting help for package 'TxDb.Hsapiens.UCSC.hg19.knownGene'
    finding HTML links ... done
    package                                 html
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (TxDb.Hsapiens.UCSC.hg19.knownGene)
* installing *source* package 'TxDb.Hsapiens.UCSC.hg38.knownGene' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
  converting help for package 'TxDb.Hsapiens.UCSC.hg38.knownGene'
    finding HTML links ... done
    package                                 html
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (TxDb.Hsapiens.UCSC.hg38.knownGene)
* installing *source* package 'org.Hs.eg.db' ...
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
```

```
** help
*** installing help indices
  converting help for package 'org.Hs.eg.db'
    finding HTML links ... done
    org.Hs.egACCNUM                         html
    org.Hs.egALIAS2EG                       html
    org.Hs.egBASE                           html
    org.Hs.egCHR                            html
    org.Hs.egCHRLENGTHS                     html
    org.Hs.egCHRLOC                         html
    org.Hs.egENSEMBL                        html
    org.Hs.egENSEMBLPROT                    html
    org.Hs.egENSEMBLTRANS                   html
    org.Hs.egENZYME                         html
    org.Hs.egGENENAME                       html
    org.Hs.egGENETYPE                       html
    org.Hs.egGO                             html
    org.Hs.egMAP                            html
    org.Hs.egMAPCOUNTS                      html
    org.Hs.egOMIM                           html
    org.Hs.egORGANISM                       html
    org.Hs.egPATH                           html
    org.Hs.egPFAM                           html
    org.Hs.egPMID                           html
    org.Hs.egPROSITE                        html
    org.Hs.egREFSEQ                         html
    org.Hs.egSYMBOL                         html
    org.Hs.egUCSCKG                         html
    org.Hs.egUNIPROT                        html
    org.Hs.eg_dbconn                        html
** building package indices
** testing if installed package can be loaded from temporary location
*** arch - i386
*** arch - x64
** testing if installed package can be loaded from final location
*** arch - i386
*** arch - x64
** testing if installed package keeps a record of temporary installation path
* DONE (org.Hs.eg.db)

The downloaded source packages are in
    ‘C:\Users\e.bordron\AppData\Local\Temp\RtmpoXryBf\downloaded_packages’
Old packages: 'cli'
Update all/some/none? [a/s/n]:
a
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/cli_3.2.0.zip'
Content type 'application/zip' length 1257150 bytes (1.2 MB)
downloaded 1.2 MB

package ‘cli’ successfully unpacked and MD5 sums checked
Warning: cannot remove prior installation of package ‘cli’
Warning: restored ‘cli’

The downloaded binary packages are in
```

```
    C:\Users\e.bordron\AppData\Local\Temp\RtmpoXryBf\downloaded_packages
Warning message:
In file.copy(savedcopy, lib, recursive = TRUE) :
  problem copying C:\Users\e.bordron\Documents\R\R-
4.1.2\library\00LOCK\cli\libs\x64\cli.dll to C:\Users\e.bordron\Documents\R\R-
4.1.2\library\cli\libs\x64\cli.dll: Permission denied
```

le mode d'emploi de rCGH est rCGH_manual.pdf dans docs. il va de pair avec rCGH.R dans scripts. à propos de l'input attendu de ce package:

```
Affymetrix SNP6.0 and cytoScanHD probeset.txt, cychp.txt, and cnchp.txt files
exported from ChAS or Affymetrix Power Tools.
rCGH also supports custom arrays, provided data complies with the expected format.
```

Les fichiers que nous avons dont le nom ressemble le plus à ça sont `2-ADREC.RC.OSCHP.chpcar`. il faudrait les comparer avec le fichier d'exemple présent dans le manuel de rCGH pour savoir si on peut les utiliser pour nos données.

J'ai aussi ajouté un tabbleur excel pour comparer les packages R. il s'agit de `review_packages_R.xlsx` dans docs/docs_I_made

arrivée à 11h -> pause 30 min , départ 18h15

24/02/2022

la convention est signée par moi + la DRH de bergonié, je l'ai envoyée à claire.

j'ai essayé d'installer des VM: workstation et virtualBox demandent le sdroits admin, donc il faut peut-être faire un ticket. J'essaie QEMU, qui possiblement marche sans droits admin mais est lent. si ça ne marche pas je fais un ticket. je continue l'excel qui compare les 3 packages. je veux me renseigner sur la question "peut-on ajuster la méthode de calcul, les coefficients, etc., dans al détermination des différents scores?" -> est-ce possible à l'aide de l'API? d'autre part, continuer à lire la doc. voir l'output de RCGH, input aussi, et l'output de EaCoN.

pour installer QEMU: j'ai choisi `C:\Users\e.bordron\Documents\qemu` comme dossier d'installation car celui par défaut renvoyait une erreur. Il s'est bien installé. Je suis ce tutoriel: `https://www.minitool.com/partition-disk/qemu-for-windows.html` pour lancer Ubuntu sur windows. edit: il est très recommandé d'installer une VM sur une partition spécialement dédiée, et je ne peux pas créer de partition. J'enverrai donc un ticket. en attendant, j'ai désinstallé QEMU. aussi: penser à installer le windows subsystem for Linux, ça peut être très pratique. voir si c'est nécessaire.

arrivée à 10h30 ; pause de 13h à 14h pour manger (30 min) + trajet à l'inrae, compté dans les heures de travail. fin à 16h40.

25/02/2022

rCGH manual (the pdf): la CGH sur array est largement utilisée en médecine, notamment pour détecter les altérations moléculaires précises. RCGH est un workflow d'analyses des données générées par cette technologie.

un workflow typique est présenté dans le document. il produit un objet rCGH qui contient:

- les infos de l'echantillon
- le dataset par sonde
- les paramètres du workflow
- les données de segmentation

un objet rCGH peut être créé à partir de readGeneric(). "CEL files have to be first read using ChAS or Affymetrix Power Tools (APT), and then exported as cychp.txt or cnchp.txt."-> peut-on faire ça?

Lire des fichiers:

les fonctions readAffySNP6() et readAffyCytoScan() permettent respectivement de lire des fichiers cychp, cnchp and probeset (.txt), exportés (par ChAS ou APT) à partir de fichiers CEL provenant des technologies SNP6.0 et CytoScanHD, respectivement. à voir si des CEL d'oncoscan peuvent faire l'affaire. D'autre part, la fonction ReadGeneric() permet de lire un custom array. il doit comporter les colonnes suivantes: ProbeName, ChrNum, ChrStart, et Log2Ratio. it creates then an *rCGH-generic* object. --> je compare mes fichiers 2-ADREC.RC.OSCHP.chpcar avec les fichiers provenant du manuel. en fait les cnchp.txt du manuel son zippés avec l'extension .bz2, que je ne peux pas ouvrir car je n'ai pas 7-zip ou un logiciel qui permet d'ouvrir ces extensions, et je ne peux pas en télécharger un non plus. Pour résoudre ça, j'ai téléchargé sur le NCBI un cnchp quelconque, que j'ai mis dans working_data sous le nom de `example_CN5.CNCHP.txt`. Sa structure est la suivante:

```
ligne 1017:
ProbeSetName    Chromosome  Position    CNState Log2Ratio   SmoothSignal    LOH
Allele Difference
CN_473963   1   61723   1   -0.869293   0.909996    nan nan
CN_473964   1   61796   1   -0.566018   0.91002 nan nan
CN_473965   1   61811   1   -0.360829   0.910024    nan nan
CN_473981   1   62908   1   -1.395133   0.910371    nan nan
CN_473982   1   62925   1   -0.353612   0.910377    nan nan
CN_497981   1   72764   1   -0.161369   0.913596    nan nan
CN_502615   1   85924   1   -0.643854   0.918221    nan nan
CN_502613   1   85986   1   -0.291606   0.918243    nan nan
CN_502614   1   86312   1   -0.487455   0.918363    nan nan
CN_502616   1   86329   1   -1.596017   0.918369    nan nan
CN_502843   1   98590   1   -0.152883   0.923038    nan nan
CN_466171   1   228694  2   -0.144197   2.308033    nan nan
CN_468414   1   229063  2   0.355558    2.308156    nan nan
CN_468412   1   229146  2   -0.523687   2.308184    nan nan
CN_468413   1   229161  2   -0.020418   2.308189    nan nan
CN_470565   1   229607  2   0.650146    2.308337    nan nan
CN_468424   1   235658  2   0.251713    2.310363    nan nan
CN_468425   1   235716  2   0.237548    2.310382    nan nan
CN_460512   1   356431  2   -0.589877   1.306899    nan nan
CN_460513   1   356530  2   -0.092504   1.3069  nan nan
```

```
SNP_A-8575125   1   564621  2   -0.237746   1.579906    1   0.840846
CN_524192   1   625458  2   -0.08702    1.813097    nan nan
CN_496034   1   707087  2   -0.065015   1.985905    nan nan
CN_500339   1   712533  2   0.209689    1.986045    nan nan
CN_502639   1   718651  2   0.096094    1.983329    nan nan
SNP_A-8709646   1   721290  2   -0.163019   1.983204    1   0.864985
SNP_A-8497791   1   740857  2   -0.080971   1.978729    1   0.940559
SNP_A-1909444   1   752566  2   -0.123229   1.978289    1   -1.062374
CN_029239   1   757457  2   0.105314    1.977908    nan nan
CN_029289   1   761356  2   -0.059604   1.977642    nan nan
...
```

à voir si je peux retrouver les mêmes colonnes dans un de mes fichiers.

télétravail + arrivée à 15h45 ; pas de pause ; 17h00, je pars.

28/02/2022

rCGH

je continue de lire la doc

adjustSignal() permet de rescale le LRR dans le cas d'affymetrix.

pour segmenter, on utilise l'algo CBS. voir notes_on_articles.md pour plus d'infos. Cette fonctionnalité peut être utilisée avec segmentCGH(). cela retourne une segmentation table.

LRR = Log2(relative ratios) la fonction EMnormalize() est utilisée pour centraliser les LRR et plotDensity() permet de visualiser cette étape et de voir quelle population a servi à centraliser les données.

note: ce package permet la parallélisation des tâches de normalisation & de segmentation de par l'utilisation du package parallel.

ce qu'on a après le workflow normal est une segmentation table.

pour convertir ça en tableau par gène:

```
byGeneTable(data)
```

additionnally, this package offers different genomic profile visualisation functions, static and interactive.

ainsi que l' `input`: un custom array doit comporter ces colonnes:

`ProbeName(probe id), ChrNum, ChrStart(The chromosomal probe locations), et Log2Ratio (amplification/deletion)` ---> ai-je ces colonnes dans un de mes documents? `present in cnchp files, present in filename.segments.txt, present in cnchp files, present in filename.segments.txt` --> partiellement dans segments.txt, partiellement dans un fichier cnchp. voir si on peut avoir un tel fichier à partir d ChAS

j'ajoute l'`output` à l'excel comparatif:

1. Segmentation table:

```
           ID chrom loc.start   loc.end num.mark seg.mean seg.med probes.Sd
estimCopy
1 CSc.Example     1    882803 249116709     1209   0.0087 -0.0504 0.9799602
2
2 CSc.Example     2     15703 242497851     1317   0.8874  0.8791 0.9901649
4
3 CSc.Example     3     62614 197683938     1100   0.8791  0.8791 0.9786349
4
4 CSc.Example     4     46691 190921709     1042  -0.0075 -0.0504 0.9883702
2
5 CSc.Example     5    113577 180579439      986   0.8502  0.8791 0.9907562
4
6 CSc.Example     6    184719 170849100     1103  -0.0105 -0.0504 1.0052332
2
```

2. byGeneTable:

```
  entrezid   symbol                       fullName cytoband chr chrStart   chrEnd
width strand Log2Ratio num.mark segNum segLength(kb) estimCopy relativeLog
genomeStart
1        1     A1BG         alpha-1-B_glycoprotein_19q13.43  19 58858172 58874214
16043      -   0.80185      231     21      58810.89         4           0
2718302494
2   503538 A1BG-AS1           A1BG antisense_RNA_1_19q13.43  19 58859117 58866549
7433      +   0.80185      231     21      58810.89         4           0
2718303439
3    29974     A1CF  APOBEC1_complementation_facto_10q11.23  10 52559169 52645435
86267      -   0.94135      751     10     135239.66         4           0
1732932312
```

pour rappel: un ratio est de la forme test/control, où control est la valeur de référence. si la valeur test est supérieure au control, le ratio va de 1 à +inf. mais si test est inférieur à control, le ratio va de 0 à 1. appliquer le log2 de ce ratio permet de rendre symétrique la répartition autour de 1.

EaCon:

le workflow typique est :

```
normalization -> segmentation +-> reporting
                              |
                              +-> copy-number estimation
```

en step-by-step:

1. Raw data processing

OS.process() will perform normalization. this step will write multiple files:

- a png: graphical representation of the normalized L2R and BAF data
- metrics of the array
- statistics ...

2. L2R & BAF Segmentation

   Segment.ff() will then perform segmentation using ASCAT, FACETS or SEQUENZA. this step will write multiple files.

   - a png: graphical representation of the segmented, centered and called L2R and BAF data
   - a png: graphical representation of BAF vs L2R of probes, by chromosome
   - various segmentation results

3. Copy-number estimation

   several parameters can be estimated in this step:

   - total and allele-specific copy-number profiles
   - global ploidy
   - sample cellularity
   - more png files The syntax is: ASCN.ff() you can also use SEQUENZA or FACETS. for more details about the output, check ASCAT R package.

arrivée à 10h55 -> 17h30, 30min pause

I move `example_CN5.CNCHP.txt` to `C:\Users\e.bordron\Desktop\CGH-scoring`.

01/03/2022

réunion aujourd'hui edit: elle est repoussée à plus tard.

je continue sur oncoscanR. J'implémente une fonction pour calculer le GI mais je me heurte (de nouveau) à ce problème:

> res = oncoscanR::workflow_oncoscan.run("C:/Users/e.bordron/Desktop/CGH-scoring/M2_internship_Bergonie/data/working_data/2-AD/2-ADREC.RC.OSCHP.segments_FULL_LOCATION.txt", "F")

```
Error in if (length(parm) == 0 || seg_start > end(parm)) { :
  missing value where TRUE/FALSE needed
In addition: Warning messages:
1: In load_chas(chas.fn, oncoscan.cov) : NAs introduced by coercion
2: In load_chas(chas.fn, oncoscan.cov) : NAs introduced by coercion
```

the cause of this problem is the values of the column Full Location containing commas.

now I use the genomic index as defined in modpathol (a pdf):

```
genomic index= A^2/C , where A is the total number of alterations (segmental gains
and losses) and C is the number of involved chromosomes.
```

I'm not sure whether I should count an arm alteration as an alteration itself. For now, let's use this: a loss or gain segment (longer than X bases) represents an alteration, where X should be changed according to data. in the case of OncoscanR results, there is no X because the length is: an arm. Anyway, I started this in OncoscanR.R

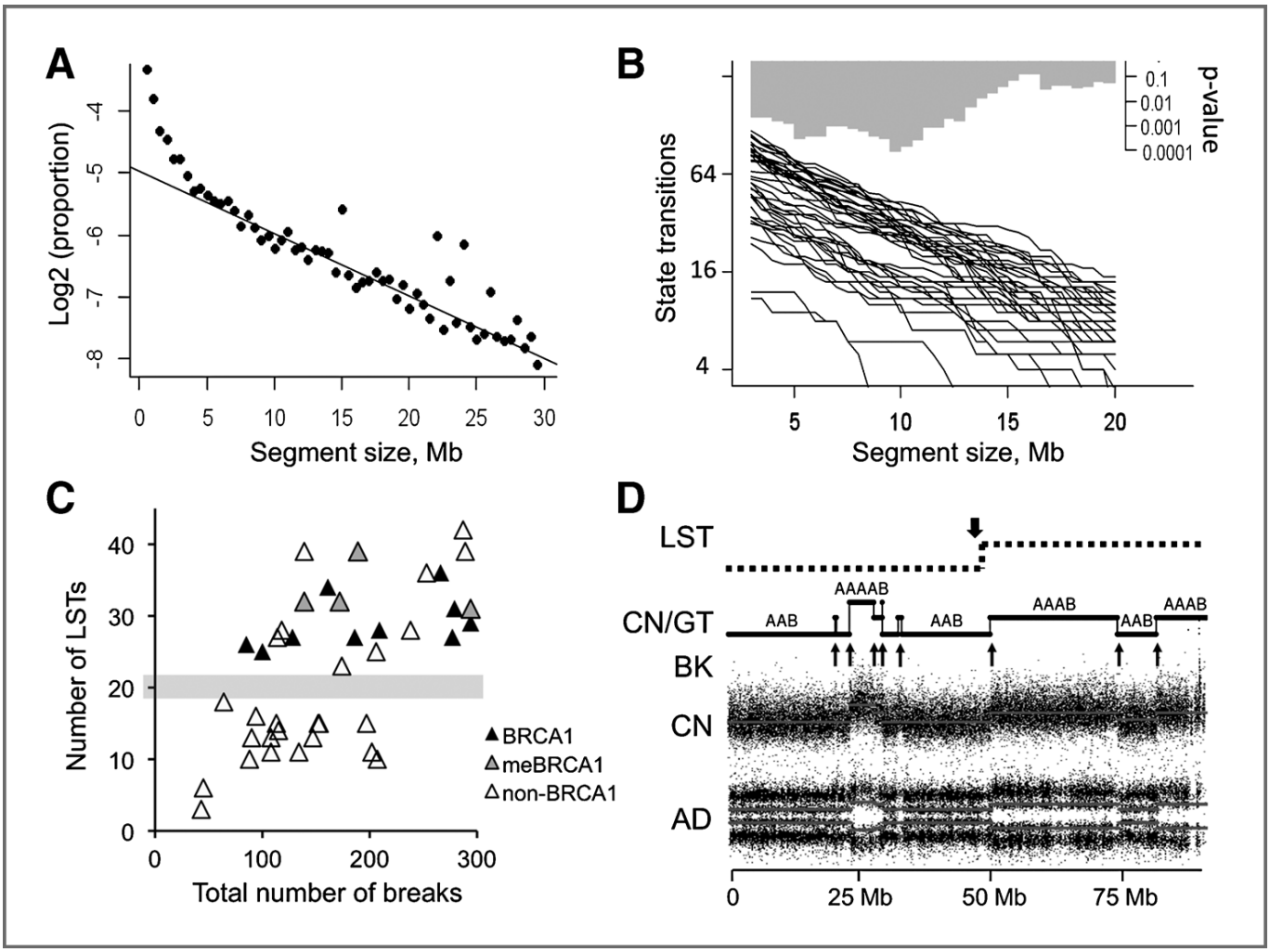
arrivée à 11h25, départ à 17:25, pas de pause.

02/03/2022

Elodie me demande de lire le document `Copy_number_aberrations_from_Affymetrix_SNP.pdf` pour savoir plusieurs choses:

- Y a-t-il une différence majeure entre ces puces là et les oncoScan?
- Les méthodes sont elles interchangeables ?

Avant ça, je lui réponds pour lui indiquer ce que sont les abréviations: LOH et LST sont 2 scores de HRD. Ils caractérisent en effet la déficience de ce pathway. LST = Large-scale State Transition. voir LST_popova.pdf un LST est un breakpoint (point de séparation entre 2 segments ayant des valeurs de CN différentes) dont les 2 segments font plus de 10 Mb. le score LST est un bon indicateur de l'état du gène BRCA1 (dont l'inactivation est souvent constatée dans le carcinome du sein). BRCA1 participe au pathway de recombination homologue (voir cahier, un procédé de réparation de l'ADN lors d'un double-strand break), et sa mutation accompagne souvent les cancers du sein ou des ovaires. Source: `Powell, S., Kachnic, L. Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. Oncogene 22, 5784–5791 (2003). https://doi.org/10.1038/sj.onc.1206678`. Une Homologous Recombination Deficiency (HRD) peut être déterminée par la mutation de BRCA1. choses nouvelles sur les LST: on aplatit les segments de moins de 3Mb (voir



), et le score LST est le nombre de LST sur tout le génome. LOH = Loss of Heterozygosity. voir LOH_abkevich.pdf Ce score correspond au nombre de segments présentant une perte d'hétérozygotie sur plus de 15 Mb. La perte d'hétérozygotie est la disparition d'un allèle sur un des deux chromosomes, supprimant du génome l'une des 2 copies de ce gène. Un lien entre ce score et une déficience du gène BRCA a été mis en évidence par les auteurs de l'article, ce qui indique que ce score est un bon indicateur de HRD. TDplus = Tandem Duplication. voir TDplus_popova.pdf Ce score est défini par le nombre de régions exprimant un gain d'une ou 2 copies (tandem duplication) et dont la taille est comprise entre 1 et 10 Mb. Le score TDplus a été lié à l'occurrence de tumeurs CDK12-deficient. Le gène CDK12 est impliqué dans la régulation de l'ARN polymérase 2, dont le dysfonctionnement va souvent de pair avec l'apparition de cancers. Source: `A ubiquitous disordered protein interaction module orchestrates transcription elongation ; 10.1126/science.abe2913`. LRR et BAF signifient respectivement Log R Ratio and B allele frequency. Sur le github d'Eacon, il est indiqué que ces valeurs sont obtenues à partir du package R rawcopy (`http://rawcopy.org/`). LRR: normalized intensity per probe relative to sample median and a reference data set BAF: estimated abundance of the B allele relative to total abundance, SNP probes only

j'essaie de déterminer ce qu'est le BAF et le LRR. en parallèle j'essaie de savoir si oncoscan appartient à Cytoscan, dans quelle mesure sont-ils différents... -> je peux télécharger les données qu'ils produisent. les comparer! -> edit: ce ne sont pas des données mais un pdf qui explique comment interpréter les données. -> j'ai trouvé cette information intéressante sur la résolution d'oncoscan (source: https://www.thermofisher.com/search/results?query=902695&persona=DocSupport&type=Product+FAQs):

```
OncoScan CNV & Oncoscan CNV Plus Assay Kit have a 50 kb-100 kb resolution in
approximately 900 cancer genes.
```

```
Outside of the cancer genes:
88% of the genome has 300 kb resolution
97% of the genome has at least 380 kb resolution
```

Je lis le document `oncoscan_sample_data_presentation.pdf` téléchargé depuis `https://www.thermofisher.com/order/catalog/product/902695?SID=srch-srp-902695` en cliquant sur `Sample Data: OncoScan&#174; FFPE Assay Kit Sample Data Presentation`. j'y retrouve le log ratio et le BAF. J'ai pris des notes à propos de ce pdf sur la feuille n°01 (écrit en violet). Je conclus de ça que le LRR et le BAF sont 2 indices très pratiques pour segmenter le génome selon le nombre de copies. ça explique qu'EaCoN fasse une segmentation à partir de ces scores.

je complète l'excel avec ces nouvelles informations. Aussi, je télécharge un pdf à lire pour mieux compprendre (peut-être) les LRR et BAF: `LRR_and_BAF.pdf`

Sur la partie qu'élodie m'a demandé de résoudre, j'ai quelques onglets d'ouverts. à re-regarder. à ce sujet, j'ai vu sur internet aujourd'hui que l'un des outputs d'oncoscan était "xxCHP.txt". peut-être que l'on peut exporter des cnchp à partir d'oncoscan, ce qui serait parfait pour l'utiliser avec R. chercher cela, si je ne trouve pas, demander à Laetitia. Chercher aussi la différence entre oncoscan et Affymetrix SNP 6.0 comme elle l'indique. je note tout ça dans todo.md. EXCEL: peut-il être envoyé?

arrivée à 10h30, 25 min pause, départ 17:50

03/03/2022

Nous sommes jeudi: ce matin, je dois aller voir Yannick pour savoir quand l'interprétation des données se fera.

je cherche aussi la différence entre oncoscan et affymetrix SNP 6.0. sur internet, je trouve le site https://www.affymetrix.com/support/developer/powertools/changelog/oschp.html .Il m'apprend des choses sur le format OSCHP, regarder ce qu'il contient sur le format CNCHP. entre autres: "The format of the OSCHP files is the HDF5 binary format." La doc d'HDF5 est ici: http://portal.hdfgroup.org/display/knowledge/HDF5+Documentation . HDFView sert à visualiser ces données. je le dl à partir de cete page: http://portal.hdfgroup.org/display/support/Download+HDFView et pas celle-ci (l'ancienne): https://support.hdfgroup.org/products/java/hdfview/index.html. OSCHP a une structure en nested tree, un peu comme du XML. aussi: "Note that the structure is an extension of the existing structure used in Affymetrix CYCHP and CNCHP file formats." J'apprends aussi qu'affymetrix encourage à utiliser Fusion SDK pour quiconque veut parser ses fichiers (la liste des fichiers affymetrix est ici: https://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/index.html ou ici: https://www.affymetrix.com/support/developer/powertools/changelog/FILE-FORMATS.html). or cette recommandation date de 2011; et depuis certains packages R se basent sur SDK pour parser les fichiers d'affy.

J'installe HDFView (à l'aide de cet installer: HDFView-3.1.3-win10_64-vs16.zip) à l'emplacement par défaut: `C:\Users\e.bordron\AppData\Local\HDF_Group\HDFView\`. Ca fonctionne bien, je peux voir le contenu d'un fichier OSCHP. voir todo.md .

Maintenant je lis `Copy_number_aberrations_from_Affymetrix_SNP.pdf`, que m'a envoyé Elodie avec la question :

```
Bonjour Elie, pourrais tu jeter un oeil à cette revue sur des outils pour
l'analyse de données Affymetrix SNP 6.0?
https://academic.oup.com/bib/article/21/1/272/5139664?login=false
je n'ai pas bien saisi si il y a une différence majeure entre ces puces là et les
oncoScan. Les méthodes sont elles interchangeables ?
```

--> je regarde ce qu'est oncoSNP, c'est l'article 19 de la biblio soit `oncoSNP.pdf`. -> je pensais qu'oncoSNP était dédié spécifiquement à oncoscan / Affymetrix SNP, mais il n'en est rien 😕 c'est pour les SNP genotyping data en général.

pour trouver une bonne fois pour toutes si SNP et oncoscan sont les mêmes puces (et si oui, je peux utiliser tous les outils de la revue `Copy_number_aberrations_from_Affymetrix_SNP.pdf` sur nos données oncoscan), je cherche leurs pages sur le site d'affymerix.

avant ça je vois dans la FAQ de rawcopy (http://rawcopy.org/FAQ) qu'oncoscan n'est pas pris en charge par ce package ("Oncoscan is a very different technology and is not supported in Rawcopy.") alors que rawcopy permet de traiter les données des microarrays CytoScan HD, CytoScan 750k et SNP 6.0 et prend des fichiers CEL en input. Pour info la dernière update de ce package date de 2019-10-13. *Cela indique qu'Oncoscan et affy SNP6.0 sont probablement différents et les méthodes sont donc non transposables.*

Je regarde les autres packages testés par l'article, voir si ils traitent les données oncoscan.

- OncoSNP: OUI

  at https://sites.google.com/site/oncosnp/frequently-asked-questions: "If you are using new array design, e.g. OncoScan, be aware that OncoSNP pre-dates these array types and is not suitably calibrated for optimal performance." -> oncosnp peut traiter oncoscan

- ASCAT: OUI

  at https://github.com/VanLoo-lab/ascat: in "Supported arrays without matched germline": "[...] AffyOncoScan [...]" Un article décrit en détail ASCAT: Allele-specific copy number analysis of tumors (ASCAT.pdf)

- GenoCNA: SÛREMENT

  at http://www.bios.unc.edu/~weisun/software/genoCN.htm and http://www.bios.unc.edu/~weisun/software/genoCN_release.htm

- GISTIC: SÛREMENT

  - input: voir GISTIC_example_seg_file.txt . The input column headers are: (1) Sample (sample name) (2) Chromosome (chromosome number) (3) Start Position (segment start position, in bases) (4) End Position (segment end position, in bases) (5) Num markers (number of markers in segment) (6) Seg.CN (log2() -1 of copy number)
  - Output: https://www.genepattern.org/doc/GISTIC_2.0/2.0.23/GISTICDocumentation_standalone.htm lire l'article GISTIC2.0facilitates_xxx.pdf pour plus d'infos.

- CGHcall: article: CGHcall_article.pdf description des fonctions du package: CGHcall_functions.pdf reference manual: CGHcall_Reference_Manual__usemewiththescript.pdf R script à utiliser avec le manuel: CGHcall.R

J'ai aussi suivi un peu l'interprétation des résultats avec Yannick. il viendra dans mon bureau me prévenir quand il fera l'interprétation avec Sabrina.

je réponds à élodie:

```
Bonsoir Elodie, j'ai regardé ce que tu m'as envoyé et certains packages sont
utilisables pour nos données, bien que les 2 méthodes (oncoscan et SNP6.0) soient
différentes.
pour OncoSNP et ASCAT, c'est sûr que les outils sont compatibles. pour les autres
ça va passer par la création de fichiers d'input, donc je devrai voir au cas par
cas. Mais avant ça il faut savoir si ils sont pertinents, donc je vais les ajouter
au tableau comparatif pour demain.
Je ne sais pas si j'aurai le temps d'avoir tous les détails pour tous les outils
mais je ferai en sorte d'avoir les inputs et outputs au minimum.
```

lire la page sur le formt de fichier CEL: https://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html . c'est intéressant arrivée à 12h00; pas de pause; départ à 17h50

04/03/2022

réunion à 14h30-15h30:

- ajouter les nouveaux packages au tableau
- demander si ils veulent d'autres indices que l'index génomique
- si oui, lesquels (LOH, LST...)

I had this problem when using git in cmd: `'git' n’est pas reconnu en tant que commande interne ou externe, un programme exécutable ou un fichier de commandes.` j'utilise le git bash maintenant, ça marche. enfin, le git bash m'indiquait le problème du disque P: , donc j'ai dû créer une variable d'environnement HOME=c: avec setx.

prochaine réunion: mardi 12h15-13h15

voir le cahier pour un récap de la réunion. Nouveaux objectifs: **tester les packages EAcon (en attente de VM), rCGH, CGHcall, oncoSNP sur 3 cas: 1 haut, un bas, un intermédiaire.**. je commence par rCGH pour pouvoir le présenter mardi. je dois lui donner un fichier en input. ---> !!!utiliser la version online pour vérifier que mes données marchent bien!!: https://fredcommo.shinyapps.io/aCGH_viewer/

J'ai vu avec Yannick comment voir quelles colonnes exporter avec ChAS.

1. ouvrir ChAS
2. charger un fichier OSCHP avec ctrl+o
3. selectionner un chromosome

4. en hait à droite, sélecitonner les colonnes à afficher
5. en haut à gauche, exporter vers un fichier texte

pour faire un input à rCGH: on peut certainement utiliser la colonne "Start Marker" en tant qu'id des sondes. *on **peut** avoir la colonne `Full Location` !*

Start Marker concerne le début d'un segment. Je peux consulter l'aide de ChAS pour savoir ce que contient chacune de ces colonnes. C'est très instructif et je peux sûrement en apprendre plus dans cette aide (500 pages env.)

arrivée à 10h30; 10 min pause, départ à 17:40

07/03/2022

faire un fichier input pour rCGH et le tester sur l'interface en ligne. l'input de rCGH est constitué d'un tableau où chaque ligne est une sonde. Or, les lignes du fichier segments.txt représentent chacune un segment altéré (LOSS, GAIN...) J'ai regardé si je pouvais trouver les intensités par sonde dans le fichier CEL. Pour cela j'ai essayé de parser le header des fichiers CEL à l'aide d'Affyio. les fonctions qui lisent directement le header ne m'apprennent rien de spécial, mais la fonction read.celfile.probeintensity.matrices() permet certainement d'obtenir ce que je veux. Cependant, un des arguments de cette fonction est `cdfInfo a list with items giving PM and MM locations for desired probesets. In same structure as returned by make.cdf.package`. Cependant `This function reads an Affymetrix chip description file (CDF) and creates an R package that when loaded has the CDF environment available for use.`

donc: demander à Laetitia & co si je peux avoir les fichiers CDF OU un tableau par sonde.

Dans rCGH_manual.pdf, on me dit de lire "Commo F, Guinney J, Ferte C, Bot B, Lefebvre C, Soria JC, and Andre F. rcgh : a comprehensive array-based genomic profile platform for precision medicine. Bioinformatics, 2015.", concernant les inputs en Custom array. Je l'enregistre sous le nom de rCGH_manual.pdf. Dans la section "3 Supported files", on m'envoie vers un lien "Supplementary Data"(un zip), qui contient Supplementary_methods.pdf. je le renomme rCGH_Supplementary_methods.pdf. Il ne m'apprend pas grand-chose de plus , mais détaille le sautres étapes et permet de mieux comprendre comment les choses se passent.

voir https://media.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cdf.html pour plus d'infos sur les CDF.

Laetitia ne sait pas comment produire un tableau avec une sonde par ligne (il serait sûrement énorme j'imagine) ni comment trouver un fichier CDF. je regarde où trouver le CDF sinon je demande à Tony Sierra.

sur http://www.aroma-project.org/FAQ/, je lis:

```
FAQ. 2007-05-24: Where can I download CDF files

A: The CDF file for a given chip type is available in so called "Library Files" at
the corresponding "Support Materials" page on the Affymetrix website. You may find
links to those Affymetrix pages via the Chip types pages: http://www.aroma-
project.org/chipTypes/
```

-> je regarde ça sur le lien d'aroma, càd les chip types pages: ils n'indiquent pas oncoscan comme un type de puce... hum. -> je regarde ça sur le site d'affymetrix. je télécharge les library files. ça met longtemps, entretemps j'ai trouvé dans la doc de rCGH ceci: https://rdrr.io/bioc/rCGH/man/readAffyOncoScan.html Sur cette page, j'apprends que rCGH peut charger un objet de classe rCGH à l'aide de la fonction readAffyOncoScan(), qui prend en argument un fichier de type `Affymetrix.tsv`. Un des scripts d'Affymetrix Power Tools (apt-copynumber-onco-ssa) permet de produire 2 fichiers 'ProbeSets,CopyNumber.tsv' and 'ProbeSets,AllelicData.tsv' dont la fusion permet de créer un tel fichier. J'en ai un exemple: `C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\docs\docs_I_found\oncoscan_affymetrix.tsv`. je l'ai construit à partir de `C:\Users\e.bordron\Documents\R\R-4.1.2\library\rCGH\extdata\oncoscan.tsv.bz2` que j'ai dézippé sur https://bz2.unzip.online/. un lien sur la page de ce script (https://www.affymetrix.com/support/developer/powertools/changelog/apt-copynumber-onco-ssa.html) indique les références qui ont permi de la construire.

Nous avons donc plusieurs possibilités:

- générer ce .tsv à l'aide du script APT
    - utiliser readAffyOncoScan() à l'aide de ce .tsv

  ou

    - utiliser readgeneric() avec un fichier input créé à partir de ce .tsv
- Si on trouve un .CDF dans les fichiers library, utiliser readGeneric() sur le fichier input créé à partir de la matrice de MM/PM donnée par la fonction read.celfile.probeintensity.matrices(.CDF).

Laquelle est la plus pertinente? il semblerait qu'oncoscan n'aie pas de CDF. je regarde après manger. ***cela implique que la meilleure façon de faire est d'utiliser readAffyOncoScan().*** J'ai trouvé un article qui indique créer des fichiers CDF pour les technologies Affymetrix. un autre package l'a utilisé. c'est un package R appelé customCDF. j'essaie de l'installer (à partir de http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp, version 25, R Package CustomCDF, Source) à l'aide de la commande suivante:

```
> install.packages("~//Users/e.bordron/Downloads/CustomCDF_1.0.5.tar.gz", repos =
NULL, type = "source")
* installing *source* package 'CustomCDF' ...
** using staged installation
** libs

*** arch - i386
Avis dans system(cmd) : 'make' not found
ERROR: compilation failed for package 'CustomCDF'
* removing 'C:/Users/e.bordron/Documents/R/R-4.1.2/library/CustomCDF'
Warning in install.packages :
  installation of package ‘C://Users/e.bordron/Downloads/CustomCDF_1.0.5.tar.gz’
had non-zero exit status
```

ça ne marche pas, j'essaie avec les binaries:

```
> install.packages("~//Users/e.bordron/Downloads/CustomCDF_1.0.5.zip", repos =
NULL, type = "win.binary")
package ‘CustomCDF’ successfully unpacked and MD5 sums checked
```

Ce package sert en fait seulement à modifier des CDF, on dirait. j'essaie d'installer non pas "R Package CustomCDF" mais "Modified 'affy' Package" à partir du binary.

```
> install.packages("~//Users/e.bordron/Downloads/affy_1.68.0.zip", repos = NULL,
type = "win.binary")
package ‘affy’ successfully unpacked and MD5 sums checked
```

je laisse tomber la piste du CDF. je vais essayer d'utiliser APT pour générer le .tsv par sonde. à ce qujet, je vois que gustave roussy (eacon) a développé un package apt.oncoscan.2.4.0 pour "processing using apt-copynumber-onco-ssa". je le teste pour voir l'output. résultat, j'ai le même problème qu'avant. c'est dans scuttle.R:

```
apt.oncoscan.process(ATChannelCel = pathToATCelFile, GCChannelCel =
pathToGCCelFile, samplename = "sample5-LD", out.dir = getwd(), force.OS =
"windows", apt.build = "na33.r2")
```

j'essaie de lancer apt-copynumber-onco-ssa. les documents dont apt a besoin pour cette étape sont dans le tableau 7 dans review_packages.xlsx . Je sais ça grâce à cette page: https://www.affymetrix.com/support/developer/powertools/changelog/VIGNETTE-OncoScan-ssa.html#libraryfiles. Je les ai téléchargés grâce à cette page: https://www.affymetrix.com/support/technical/byproduct.affx?product=oncoscan_assay_kits. Je ne peux pas installer APT. j'ai utilisé ce lien: https://www.thermofisher.cn/cn/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html?adobe_mc=MCMID%7C00326299451309961883732216901771516756%7CMCAID%3D3107E39F4245E90B-6000164143BD1666%7CMCORGID%3D5B135A0C5370E6B40A490D44%40AdobeOrg%7CTS=1614293705 Je télécharge ainsi un installer qui se trouve ici: `C:\Users\e.bordron\Desktop\CGH-scoring\apt_2_11_4_windows_installer-win64\apt_2_11_4_windows_installer-win64`

j'essaie CGHcall en attendant d'avoir la machine virtuelle qui me permettrait de lancer ça. regarder comment les auteurs de l'article qui a comparé les 6 outils pour les SNP ont utilisé ce package (ils disent à un moment: pour tel pkg, on a calculé tel truc, puis on a fait ça pour qu'il soit comparable aux autres.)

réunion starleaf pour demain, lien visio: https://meet.starleaf.com/4597475549/app

arrivée à 9h25; 20 min pause; départ à 18:35

08/03/2022

les fichiers library d'oncoscan requis pour lancer apt-copynumber-onco-ssa font plus de 50MB. je les retire du dépot git, mais ils restent dans CGH-scoring. les auteurs de la comparaison de 6 outils disent avoir créé cghcall*, mais ne le rendent disponible nulle part. Un fichier Tex constitue leur "additional material", par contre. CGHcall est concu pour du aCGH. Il utilise les informations de breakpoint (par l'algo CBS, typiquement) et classifie le LRR entre ref et tumeur en 5 états.

résumé de la réunion d'aujourd'hui: mon script pour calculer le GI à partir d'OncoscanR a montré que ça marche bien pour certains cas. je l'applique à tous mes résultats. donc je continue sur le script oncoscanR.R:

- nombre de chromosomes automatique
- process all files in one script

je passe tous les echantillons dans ChAS pour obtenir les segments.txt avec la colonne Full Location. ces nouveaux fichiers sont de la forme ?-XX.OSCHP.segments.txt où ?=un nombre et XX=deux lettres. par exemple: 3-ES.OSCHP.segments.txt problème: quand je passe les données de 2-AD dans Chas ,je n'obtiens pas du tout ce que laetitia m'avait envoyé. il y a plus de lignes. je vais voir ça avec elle après avoir regardé pour un autre échantillon. j'enregistre 11-BG.OSCHP.segments_clean_fl.txt. clean-fl doit être ajouté pour les fichiers que je crée ainsi. je le compare avec l'original dans oncoscanR (je commit maintenant, j'inclus le terme "findme5186841132" dans le message de commit). Je vais voir laetitia pour lui en parler. comme support, je lui envoie un excel: "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\working_data\2-AD\generer_des_fichiers_segments.xlsx" Pb résolu, j'utilisais un OSCHP "nettoyé". Afin de travailler sur les données brutes, je vais utiliser les CEL que j'ai reçu pour créer les OSCHP bruts à l'aide de ChAS. Pour cela:

1. ouvrir ChAS
2. Analysis
3. Perform Analysis Setup
4. use a batch file to tell ChAS which CEL files to take and which ouptut name they should use
5. specify "C:\Users\e.bordron\Desktop\CGH-scoring\interact_with_CHAS" for all paths
6. click Submit. I wasn't able to provide the AT CEL file for sample 1. it was present but ChAS did not offer me the possibility to click it.

ChAS a bien généré les OSCHP, j'ai vérifié, 0 erreurs et 0 warnings. ces fichiers sont dans "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\data\working_data\from_CEL". J'ai généré les fichiers segments.txt à partir des OSCHP, ils sont au même endroit. D'autre part, les fichiers segments ont parfois une ou deux lignes seulement. je pense que c'est normal, ça doit être dû au fait qu'ils ne sont pas recentrés, et par conséquent ils ne repèrent que peu d'altérations. le GI sera naturellement moins précis, par rapport aux données recentrées par laetitia. Les fonctions sont faites dans oncoscanR.R. j'obtiens ainsi un GI de 80 pour l'echantillon 2 par exemple. c'est élevé, je me demande si je fais bien de compter 14q et 14p comme deux altérations. j'implémente une deuxième méthode de calcul où si les 2 bras d'un chromosome sont altérés, une seule altération est comptée (j'implémente surtout la possibilité de choisir l'une ou l'autre en argument de la fonction principale).

Je vais devoir continuer le script et le faire tourner sur tous les échantillons. aussi faire la même chose sur un autre package et envoyer un mail à jennifer pour la VM.

arrivée à 9h55; 15 min pause; départ à 18h45

09/03/2022

obj: continuer le script. laetitia m'a envoyé "tableau rendu résultats ONCOSCAN.xlsx". je le mets dans "C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\docs". le remplir avec les résultats d'oncoscanR. le script produit les computations d'oncoscanR automatiquement pour tous les échantillons. output: une liste de longueur n où chaque élément est un output d'oncoscanR (une liste R). la prochaine étape est de traiter ces résultats en routine avec la fonction calcGI. calcGI fonctionne et la liste de résultats aussi. prochain obj: traiter en routine tous les résultats, créer ainsi une colonne "GI_oncoscanR" dans un dataframe qui contient déjà une colonne "sample_id" et l'exporter en fichier texte.

pour split un string en utilisant '.' comme séparateur , ne pas oublier d'échapper le point avec deux antislash, car le point seul est un regex qui dit "n'importe quel caractère". exemple d'application: str_split(filename, '\.'). voir oncoscanR.R .

oncoscanR.R est clean, a un main et tout ce qu'il faut.

arrivée à 9h50; 20 min pause; départ à 18h10

10/03/2022

laetitia m'a fait un retour sur les GI calculés par oncoscanR: retour_de_laetitia__calcul_GI_oncoscanR.xlsx j'ai relancé jennifer pour la VM.

fatal: the remote end hung up unexpectedly github a de nouveau un pb. de PATH. je me rends compte que depuis le début, j'ajoutais git au path des variables user, pas le path système. je ne peux pas modifier ce dernier à la main. j'ai réinstallé git et coché l'option qui dit "git sera ajouté au PATH, mais les fonctions find et truc natives de windows seront override par celles de git. ne choisissez cette option que si vous comprenez les implications". git est installé dans `C:\Users\e.bordron\Documents\Git_custom_install` et il fonctionne avec la console cmd.

je continue les recherches sur CGHCall. dans le script démo, un dataset "wilting" est utilisé. il vient de l'article `Increased gene copy numbers at chromosome 20q are frequent in both squamous cell carcinomas and adenocarcinomas of the cervix (CGHcall_wilting_dataset.pdf)`. je regarde à quoi correspondent les colonnes dans R. Dur à dire, j'ai l'impression que les colones SSC sont des échantillons... je lis l'article de CGHcall. je l'ai en pdf.

voir todo.md

voir la feuille sur mon bureau pour compléter les infos sur CGHcall. arrivée à 10:05; pas de pause; départ 17h55.

11/03/2022

je remplis les infos sur CGHcall à l'aide de ma prise de notes papier.

En ce qui concerne le dataset, la forme est la suivante:

```
    BAC.clone CHROMOSOME START_POS END_POS     AdCA10     SCC27      SCC32
SCC36     SCC39
1   CTB-14E10          1    941583 1144379         NA        NA         NA
```

```
NA        NA
2 RP11-465B22          1    986153 1117131 -0.1714960 0.5638540 -0.3082376
0.1353897 0.2322698
3  RP4-785P20          1   3214521 3355092 -0.1227932 0.7255571 -0.3397409
0.1236178 0.2652958
4   RP1-37J18          1   4476787 4608114  0.3007097 0.6012696 -0.2582796
0.1978820 0.2252346
5   RP11-49J3          1   5866139 5966440         NA        NA         NA
NA        NA
6   RP3-438L4          1   7059893 7146992  0.3032238 0.5992452 -0.2121089
0.2361994 0.3708779
```

BAC.clone correspond à une sonde, CHROMOSOME est explicite, START_POS et END_pos définissent les coordonnées de la sonde

Important: on peut exporter un fichier probeset.txt, donc où une ligne correspond à une sonde, à l'aide de ChAS. Pour cela, suivre la procédure du manuel d'aide de ChAS (disponible à `C:\Users\e.bordron\Desktop\CGH-scoring\M2_internship_Bergonie\docs\docs_I_found\ChASRUOHelp.pdf`) page 91, mais au lieu de cliquer sur "Export QC table" à l'étape 4, cliquer sur "generate report" puis sur "export probe level data".

arrivée à 9h45; départ à 17:40 ; pas de pause.

Lundi 14/03/2022

Elodie a envoyé ça sur Slack:

L'objectif: développer une expertise en bioinformatique et acquérir les bonnes pratiques

- Comprendre la question biologique
- Faire un état de l'art
    - sur la question biologique (bibliographique accompagné de discussions avec les biologistes si possible)
    - Sur les méthodes publiées qui ont concerné des questions similaires (bibliographique, c'est vous les experts)
        - Y en a-t-il des intéressantes ? Oui: Pourquoi ? Non: Pourquoi, que puis-je proposer ?
- Comprendre les données
    - Quelle technologie utilisée pour les produire?
    - Quels sont les grandes étapes des expériences ?
    - Quels biais inhérents à la technologie/expérience ?
    - Quel design expérimental ?
    - Que représentent les données
        - Quel type de mesures ? Continues/discrètes/catégorielles
        - Quels sont les paramètres à prendre en compte ?
        - Quelle(s) normalisation(s) appliquer ?
        - Comment les biais expérimentaux se reflètent dans les données ?
- Dans le cadre de comparaison de méthodes pour un type de données
    - Comprendre les méthodes

        - Dans quel cadre la méthode a été développée, comment se compare-t-elle aux méthodes alternatives ?
        - Comment sont processées les données ? (eg normalisation, transformation)
        - Comment sont représentés les résultats ?
    - Quelles méthodes semblent les plus appropriées pour répondre à la question
        - Facilité de mise en oeuvre
        - Output adéquat
        - Spécificité en terme d'approche (processing, normalisation et transformation)
        - Spécificité en terme d'extraction de résultats (quelles métriques, quels graphiques etc)
- Dans le cadre d'utilisation de données publiques
    - Les données sont brutes : appliquer les questions du paragraphe précédent
    - Les données sont traitées :
        - Quels traitements ont été appliqués ?
        - Quels outils ?
        - Pour répondre à quelle question ?
        - Est ce que ça correspond à la question à la quelle on veut répondre ?
        - Y a-t-il d'autres traitements à faire ?
- Bien identifier le but de chaque analyse / chaque ligne de code
- Penser au rendu des résultats
    - Quelle illustration est la plus adaptée pour faire passer mon message (tableau, plot en tout genre) ?
    - De quoi à besoin le lecteur pour comprendre ce que j'ai fait ?

Garder ça en tête; essayer de répondre à tout.

Finir de noter les infos sur CGHcall; noter les nouveautés de ce package sur le tableau comparatif

Je me renseigne également sur le package DNAcopy. les documents sont téléchargés. un package Bioconductor appelé snapCGH utilise un modele de Markov pour la segmentation. à voir si c'est intéressant.

utilisation de log_book.md: n'y mettre que des questions. pour répondre aux questions, ne mettre que des références vers les autres fichiers texte.

pour modifier le syntax highlighting: "From the official Doc: To tune the editor's syntax highlighting colors, use editor.tokenColorCustomizations in your user settings settings.json file" ex:

```
"editor.tokenColorCustomizations": {"textMateRules": [{
        "scope": "keyword.control.ref.latex",
        "settings": {
            "foreground": "#FF0000"
        }
    }]}
```

Et pour trouver les scopes, faire `ctrl+maj+P` puis `Developer: Inspect Editor Tokens and Scopes` . Attention, il n'y a pas de retour à la ligne dans le nom des scopes, parfois il y a juste plusieurs scopes pour un même élément.

J'ai ajouté DNAcopy dans notes_on_articles.

arrivée à 9h20; départ à 19h00; 40 min pause

Mardi 15/03/2022

Aujourd'hui, réunion à 12h15 ici: https://meet.starleaf.com/4828136196/app

faire des plots avec CGHcall pour la réunion, expliquer l'intérêt de la segmentation, faire tourner peut-être? non pas besoin, on a le sample data de CGHcall. Parler aussi du fait que oncoscanR donne de mauvais résultats sans recentrage/normalisation, je dois regarder si c'est possible de lui en faire faire une. essayer de faire ça avant la réunion. mettre à jour le tableau comparatif. Plein d'infos ajoutées dans le compte-rendu de réunion + le cahier. je teste VMware. je crée une nouvelle VM, je sélectionne linux.iso, Ubuntu, je l'appelle Ubuntu_bergo, elle se trouve à `C:\Users\e.bordron\Documents\Virtual Machines\Ubuntu_bergo`. ça n'a pas marché, il fallait donner un .iso. j'ai dl l'iso ubuntu, je l'ai utilisé, elle a détecté que c'était Ubuntu. Elle fait donc une Easy install: Full name: elie bordron User name: ebor mdp: 7Cha

J'appelle la machine `Ubuntu_64-bit_bergo_v2`. elle est dans `C:\Users\e.bordron\Documents\Virtual Machines\Ubuntu_64-bit_bergo_v2`. Je lui donne 21GB d'espace libre. je coche "split virtual disk into multiple files"

J'ai installé R sur la VM ubuntu. J'ai aussi fait un script R pour afficher les courbes normales relatives aux segments called en -1, 0 et 1, afin de les superposer à la courbe générale. cela permet de mieux comprendre le fonctionnement du mixture model pour le présenter en slides. le script est CGHcall.R, à compléter et surtout à *valider*

$all markdown rules$

arrivée à 10h05; départ à 19H45 ; 30 minutes de pause

Mercredi 16/03/2022

continuer le script: valider ce qu'il fait. Fait. le script fonctionne et génère 2 plots intéressants

avant de partir:

- ~~avoir un script lisible et facile à retrouver ce qu'il fait.~~
- noter ce que j'ai compris sur CGHcall: comment fonctionnent les mixture models, et le fait que pour CGHcall, les points utilisés sont les segments générés par l'étape de segmentation.
- ~~envoyer les 2 plots à élodie/les stocker qqpart.~~
- parler à élodie aussi du fonctionnement du package, avec segmentation couplée au MM? Pas besoin, je le ferai en lui montrant les slides.
- faire une slide basique par étape? et continuer le ppt demain

arrivée à 10h00; départ à 18:15; 30 min pause

Jeudi 17/03/2022

Jeudi 10h: visio avec Slim Karkar

puis-je faire un jour de télétravail?

1. ~~lire ma convention~~
2. ~~demander à caroline en précisant la date.~~

fonctionnement_des_articles.md dcrit le pipeline de CGHcall. je me sers du pipeline pour faire 1 slide par étape importante. Faire ceci pour les 4 outils.

J'installe aussi EaCoN sur la vm ubuntu:

INSTALLATION

CORE

```
install.packages('devtools')
```

```
devtools::install_github("Crick-CancerGenomics/ascat/ASCAT")
devtools::install_github("mskcc/facets")
```

l'installation n'a pas marché pour FACETS; j'utilisais la version 3.6.3 de R. J'ai installé la version 4.1.3 de R à l'aide de cette page web: https://askubuntu.com/questions/1341755/installing-r-4-0-2-on-ubuntu-20-04 qui cite cette dernière: https://cran.rstudio.com/bin/linux/ubuntu/

j'installe les paquets nécessaires. en installant devtools, j'obtiens cette erreur:

> install.packages("devtools", dependencies = TRUE)

```
Error: package ‘usethis’ was installed before R 4.0.0: please re-install it
Execution halted
```

Bien que l'argument dependencies est censé installer les dépendances d'un package, il est coincé lorsqu'une dépendance est installée dans une mauvaise version. j'ai trouvé cette page web: https://stackoverflow.com/questions/58892908/when-installing-an-r-package-automatically-reinstall-dependencies-when-needed Qui indique que pak::pkg_install() serait une bonne alternative. ça a effectivement marché après avoir dû installer manuellement 2-3 packages.

voir https://stackoverflow.com/questions/71404215/error-in-install-github-system-command-rcmd-exe-failed-exit-status-1-stdou ; une partie de l'installation ne se fait pas. demander à jennifer si elle arrive à télécharger ça sur son PC bergonié.

arrivée à 9h35; départ à 18:20; 25 min pause;

Vendredi 18/03/2022

Ceux qui ont résolu le problème que j'ai pour devtools l'ont fait en réinstallant l'OS. je recrée la machine virtuelle de a à z. Fait. j'ai aussi changé le clavier (raccourci pour passer d'un keyboard layout à un autre: super + espace). J'installe R version 4.1.3 directement; peut-être qu'avoir 2 versions installées causait mon problème.

J'ai vérifié la clé, elle fonctionne. c'était le cas aussi pour la première installation. J'installe aussi Rstudio avec le .deb que j'ai téléchargé.
R avec Rstudio s'ouvre.
des messages d'erreur m'indiquent d'installer des packages avec apt:
"-------------------------- [ANTICONF] --------------------------------"
-> je les installe. j'ai aussi plusieurs fois eu cette ligne: `/usr/lib/R/bin/config: 1: eval: make: not found` après avoir installé les anticonf, je relance la ligne de commande pour installer devtools. il me les redemande; je ferme et rouvre R.
Toujours le même problème, mais en fait le message me dit quoi faire:

```
--------------------------- [ANTICONF] -------------------------------
Configuration failed to find the fontconfig freetype2 library. Try installing:
 * deb: libfontconfig1-dev (Debian, Ubuntu, etc)
 * rpm: fontconfig-devel (Fedora, EPEL)
 * csw: fontconfig_dev (Solaris)
 * brew: freetype (OSX)
If fontconfig freetype2 is already installed, check that 'pkg-config' is in your
PATH and PKG_CONFIG_PATH contains a fontconfig freetype2.pc file. If pkg-config
is unavailable you can set INCLUDE_DIR and LIB_DIR manually via:
R CMD INSTALL --configure-vars='INCLUDE_DIR=... LIB_DIR=...'
```

0. j'ai appris que PKG_CONFIG_PATH était l'endroit où pkg-config allait chercher les fichiers .pc (source: `https://askubuntu.com/questions/210210/pkg-config-path-environment-variable`)
1. j'ai installé locate.
2. j'ai cherché le .pc que R me demande d'ajouter au path de pkg-config à l'aide de la commande `locate freetype2.pc`. output: `/usr/lib/x86_64-linux-gnu/pkgconfig/freetype2.pc`.
3. j'ai ajouté le dossier qui contient ce .pc à PKG_CONFIG_PATH: `export PKG_CONFIG_PATH="/usr/lib/x86_64-linux-gnu/pkgconfig"`. Je relance ensuite install.packages("devtools", dependencies = T). Même output. le message précise qu'il faut donner le path vers le *fichier* .pc, pas le dossier qui le contient. j'essaie ça. avant ça, je relance R et je relance la commande. même output. je relance après avoir changé le path pour .pc . même output. je lance install.packages('devtools'), pour voir. j'obtiens ceci:

```
------------------------ ANTICONF ERROR ---------------------------
Configuration failed because libcurl was not found. Try installing:
 * deb: libcurl4-openssl-dev (Debian, Ubuntu, etc)
 * rpm: libcurl-devel (Fedora, CentOS, RHEL)
 * csw: libcurl_dev (Solaris)
If libcurl is already installed, check that 'pkg-config' is in your
PATH and PKG_CONFIG_PATH contains a libcurl.pc file. If pkg-config
is unavailable you can set INCLUDE_DIR and LIB_DIR manually via:
R CMD INSTALL --configure-vars='INCLUDE_DIR=... LIB_DIR=...'
-------------------------------------------------------------------
```

Bien que PKG_CONFIG_PATH contienne les bons chemins:

> echo $PKG_CONFIG_PATH

```
/usr/lib/x86_64-linux-gnu/pkgconfig/freetype2.pc:/usr/lib/x86_64-linux-
gnu/pkgconfig/libcurl.pc:/usr/lib/x86_64-linux-gnu/pkgconfig
```

j'ai toujours ce message d'erreur:

j'installe pak (utiliser `install.packages("pak", type = "source")` !). je fais `pak::pkg_install("devtools")` -> `Error in load_private_package("glue") : Cannot load glue from the private library` je suis les informations de cette page : `https://stackoverflow.com/questions/71246072 error-in-load-private-packageglue-cannot-load-glue-from-the-private-librar` et je le réinstalle avec `install.packages("pak", type = "source")` j'ai toujours la même erreur pour glue. je réinstalle R et je spécifie, dans la fonction install.packages(), l'argument lib qui indique où installer les librairies. -> je choisis un endroit où pkg-config ira chercher naturellement.
Je peux aussi lancer cette commande: `R -e '.libPaths()'` pour trouver les fichiers lib installés. ainsi je les supprime et les réinstalle.

```
les paths sont les suivants:
/home/ebor/R
/usr/local/lib/R
/usr/lib/R
```

je désinstalle R de nouveau, puis je supprime manuellement ces trois dossiers. seulement `/home/ebor/R` était encore présent. je réinstalle (j'ai juste besoin de taper `apt install --no-install-recommends r-base`) maintenant:

1. à l'aide de `pkg-config --variable pc_path pkg-config` je détermine dans quels dossiers pkg cherche les librairies installées:

   ```
   /usr/local/lib/x86_64-linux-gnu/pkgconfig
   /usr/local/lib/pkgconfig
   /usr/local/share/pkgconfig
   /usr/lib/x86_64-linux-gnu/pkgconfig
   /usr/lib/pkgconfig
   /usr/share/pkgconfig
   ```

2. je spécifie `/usr/lib/pkgconfig` pour chaque install.packages():

> install.packages("devtools", lib="/usr/lib/pkgconfig")

```
Warning in install.packages :
  'lib = "/usr/lib/pkgconfig"' is not writable
```

je spécifie `/usr/share/pkgconfig` pour chaque install.packages(): même erreur. je rends le dossier `/usr/lib/pkgconfig` accessible en écriture. la ocmmande se lance. je remarque que des messages d'erreur indiquaient que gcc et g++ n'étaient pas installés, c'est fait maintenant. cependant, libcurl a été installé dans l'ancien repo, le par défaut.

> pak::pkg_install("devtools")

```
✓ Loading metadata database ... done

→ Will install 71 packages.
→ All 71 packages (22.99 MB) are cached.
+ askpass       1.1    [bld][cmp]
[...]
+ xopen         1.0.0  [bld]
+ yaml          2.3.5  [bld][cmp]
+ zip           2.2.0  [bld][cmp]
ℹ No downloads are needed, 71 pkgs (22.99 MB) are cached
ℹ Building brew 1.0-7
ℹ Building brio 1.1.3
✓ Built brew 1.0-7 (2s)
[...]
ℹ Building devtools 2.4.3
✓ Built devtools 2.4.3 (3.8s)
✓ Installed devtools 2.4.3  (40ms)
✓ 1 pkg + 70 deps: added 71 [4m 54.8s]  pak::pkg_install("devtools")
✓ Loading metadata database ... done

→ Will install 71 packages.
→ All 71 packages (22.99 MB) are cached.
+ askpass       1.1    [bld][cmp]
[...]
+ xfun          0.30   [bld][cmp]
+ xml2          1.3.3  [bld][cmp]
+ xopen         1.0.0  [bld]
+ yaml          2.3.5  [bld][cmp]
+ zip           2.2.0  [bld][cmp]
ℹ No downloads are needed, 71 pkgs (22.99 MB) are cached
ℹ Building brew 1.0-7
ℹ Building brio 1.1.3
✓ Built brew 1.0-7 (2s)
ℹ Building clipr 0.8.0
✓ Built brio 1.1.3 (2.8s)
[...]
✓ Built testthat 3.1.2 (45s)
✓ Installed testthat 3.1.2  (145ms)
ℹ Building devtools 2.4.3
✓ Built devtools 2.4.3 (3.8s)
✓ Installed devtools 2.4.3  (40ms)
✓ 1 pkg + 70 deps: added 71 [4m 54.8s]
```

> install.packages("BiocManager) BiocManager::install("GenomicRanges") BiocManager::install("GenomicRanges") devtools::install_github("Crick-CancerGenomics/ascat/ASCAT") devtools::install_github("mskcc/facets") -> sur cette commande, Rstudio plante. je la lance depuis une console R lancée depuis un terminal

```
Downloading GitHub repo mskcc/facets@HEAD
Downloading GitHub repo veseshan/pctGCdata@HEAD
   checking for file ‘/tmp/RtmpK79XIP/remotes6ee54bf2cff2/veseshan-pctGCdata-
d2d✔  checking for file ‘/tmp/RtmpK79XIP/remotes6ee54bf2cff2/veseshan-pctGCdata-
d2d4faf/DESCRIPTION’
─  preparing ‘pctGCdata’:
✔  checking DESCRIPTION meta-information ...
─  installing the package to process help pages
         ----------------------------------
─  installing *source* package ‘pctGCdata’ ...
   ** using staged installation
   ** R
   ** data
   *** moving datasets to lazyload DB
   Killed
         ----------------------------------
   ERROR: package installation failed
Error: Failed to install 'facets' from GitHub:
  Failed to install 'pctGCdata' from GitHub:
  System command 'R' failed, exit status: 1, stdout & stderr were printed
```

j'essaie d'installer directement pctGCdata avec les 2 façons disponibles mais ça ne marche pas. Demander à Jennifer si elle arrive à télécharger facets sur son ubuntu.

Je recrée une VM, réinstalle R version 4.1.2 (la même que sur windows), et retente. la vm s'appelle v3, je la construis avec les mêmes paramètres que les 2 premières. je ne supprime pas la v2, au cas où.

sur v3, je fais ceci:

```
# update indices
apt update -qq
# install two helper packages we need
apt install --no-install-recommends software-properties-common dirmngr
# add the signing key (by Michael Rutter) for these repos
# To verify key, run gpg --show-keys /etc/apt/trusted.gpg.d/cran_ubuntu_key.asc
# Fingerprint: 298A3A825C0D65DFD57CBB651716619E084DAB9
wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc | sudo
tee -a /etc/apt/trusted.gpg.d/cran_ubuntu_key.asc
# add the R 4.0 repo from CRAN -- adjust 'focal' to 'groovy' or 'bionic' as needed
add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu $(lsb_release
-cs)-cran40/"
```

puis `sudo apt policy r-base`.
output:

```
r-base:
  Installed: (none)
  Candidate: 4.1.3-1.2004.0
  Version table:
     4.1.3-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.1.2-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.1.1-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.1.0-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.5-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.4-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.3-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.2-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.1-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.0-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     3.6.3-2 500
        500 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 Packages
        500 http://us.archive.ubuntu.com/ubuntu focal/universe i386 Packages
```

j'installe donc la 4.1.2 avec:
`sudo apt-get install r-base=4.1.2-1.2004.0`

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
Some packages could not be installed. This may mean that you have
requested an impossible situation or if you are using the unstable
distribution that some required packages have not yet been created
or been moved out of Incoming.
The following information may help to resolve the situation:

The following packages have unmet dependencies:
 r-base : Depends: r-recommended (= 4.1.2-1.2004.0) but 4.1.3-1.2004.0 is to be
installed
          Recommends: r-base-html but it is not going to be installed
          Recommends: r-doc-html but it is not going to be installed
E: Unable to correct problems, you have held broken packages.
```

Voir les réponses à `https://askubuntu.com/questions/564282/apt-get-unmet-dependencies-but-it-is-not-going-to-be-installed`. Si je ne peux rien faire avec, demandr à jennifer.

J'ai utilisé ffmpeg sur windows 10 pour faire un gif. la commande utilisée est `%ffmpeg% -framerate 0.5 -i %03d.png output.gif` en étant dans le dossier qui contient les images. ces dernières doivent s'appeler 001.png, 002.png, etc. car `%03d` correspond au nombre de digits qui composent le nom de l'image (sans l'extension .png).

arrivée à 10h05; départ à 17:20; 1h pause

Luni 21/03/2022

installer Rstudio sur ubuntu: télécharger le .deb sur le site de Rstudio, puis `sudo apt-get install /home/Downloads/filename.deb`. j'ai utilisé aptitude pour sélectionner la version de R que je veux isntaller (4.1.2) et résoudre le conflit, mais aptitude a installé quand même 4.1.3. j'essaie d'installer pak et de lancer les installs au cas où. glue problem. besoin de réinstall R. je désinstalle R et j'essaie l'option -f pour forcer l'install de 4.1.2. (avec la commande `sudo apt-get install r-base=4.1.2-1.2004.0`). toujours le glue problem. je désinstalle vraiment R en cherchant les paths des libs de R, voir la commande `R -e '.libPaths()'`. elle doit être lancée avant que R ne soit désinstallé, bien sur. l'output:

```
"/home/ebor/R/x86_64-pc-linux-gnu-library/4.1"
"/usr/local/lib/R/site-library"
"/usr/lib/R/site-library"
"/usr/lib/R/library"
```

Les trois derniers sont supprimés quand R est désinstallé. R est complètement désinstallé, je fais alors `sudo apt-get install r-base=4.1.2-1.2004.0 -f`. output:

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
Some packages could not be installed. This may mean that you have
requested an impossible situation or if you are using the unstable
distribution that some required packages have not yet been created
or been moved out of Incoming.
The following information may help to resolve the situation:

The following packages have unmet dependencies:
 r-base : Depends: r-recommended (= 4.1.2-1.2004.0) but 4.1.3-1.2004.0 is to be
installed
          Recommends: r-base-html but it is not going to be installed
          Recommends: r-doc-html but it is not going to be installed
E: Unable to correct problems, you have held broken packages.
```

R n'a pas été installé. J'essaie la version aptitude: `sudo aptitude install r-base=4.1.2-1.2004.0 -f` option2, "installer r-recommended 4.1.2" -> installe R 4.1.3 j'installe pak avec `install.packages("pak",`

`type = "source")` afin d'éviter le probleme glue le problème glue survient quand même. peut-être que préciser l'argument lib="/usr/lib/pkgconfig" pourrait mieux marcher. je désinstalle pak et je réessaye avec cet argument. Fait, ça télécharge un tar.gz qu'il faut ouvrir pour installer le package. quand je le fais, j'ai l'erreur de glue.

à essayer: installer d'abord r-recommended 4.1.2, puis r-base 4.1.2 .je désinstalle R et je fais ça. juste en installant recommended, r 4.1.3 s'installe.

Je cherche une autre façon d'installer la version 4.1.2 de R, comme ça j'aurai au moins la même version pour comparer entre windows et ubuntu pour EaCoN.

je lis *https://askubuntu.com/questions/1373827/problems-installing-latest-version-of-r-in-ubuntu-20-04-lts*. J'ouvre la vm v2.

> apt-cache policy r-base

```
r-base:
  Installed: (none)
  Candidate: 3.6.3-2
  Version table:
     3.6.3-2 500
        500 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 Packages
        500 http://us.archive.ubuntu.com/ubuntu focal/universe i386 Packages
```

En suivant la première réponse de la page web, je fais non pas ça: `sudo add-apt-repository 'deb https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/'` mais ça: `add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu $(lsb_release -cs)-cran40/"` ensuite:

```
apt-cache policy r-base
r-base:
  Installed: (none)
  Candidate: 4.1.3-1.2004.0
  Version table:
     4.1.3-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.1.2-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.1.1-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.1.0-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.5-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.4-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.3-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.2-1.2004.0 500
        500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
     4.0.1-1.2004.0 500
```

```
          500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
      4.0.0-1.2004.0 500
          500 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ Packages
      3.6.3-2 500
          500 http://us.archive.ubuntu.com/ubuntu focal/universe amd64 Packages
          500 http://us.archive.ubuntu.com/ubuntu focal/universe i386 Packages
```

je suis l'installation du poseur de question:

> sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys E298A3A825C0D65DFD57CBB651716619E084DAB9

```
Executing: /tmp/apt-key-gpghome.h9ej4r0hgd/gpg.1.sh --keyserver
keyserver.ubuntu.com --recv-keys E298A3A825C0D65DFD57CBB651716619E084DAB9
gpg: key 51716619E084DAB9: "Michael Rutter <marutter@gmail.com>" 1 new signature
gpg: Total number processed: 1
gpg:         new signatures: 1
```

puis:

> sudo apt update && sudo apt upgrade puis: sudo apt install r-base -> output: a lot of text. on dirait que pak était déjà installé puisqu'en ouvrant R j'ai pu faire `pak::pkg_install("devtools")` directement. output:

```
✓ Updated metadata database: 2.57 MB in 3 files.
✓ Updating metadata database ... done

ℹ No downloads are needed
✓ 1 pkg + 69 deps: kept 70 [10.3s]
```

je viens d'envoyer un mail à julie, jennifer et elodie pour leur demander si l'installation de facets plante sur leurs linux, aussi. Elodie a répondu: facets n'est pas obligatoire. lire l'article eacon pour compléter le résumé que j'en ai fait à l'instant dans fonctionnement_des_outils.md. Edit: il n'y en a pas. mais lire l'article d'ASCAT par contre. aussi: regarder comment chaque fonction d'EaCoN fait ses opéations: quelles fonctions de quels packages appelle-t-elle? comment fonctionnent ces dernières?

arrivée à 10h35; départ à 18:00; 30 min pause

Mardi 22/03/2022

réunion aujourd'hui.

- avoir le pipeline décrit pour les 4 packages. jusque-là, j'ai CGHcall et EaCoN.
- noter ce que j'ai surligné dans ASCAT_article.pdf. 2 paramètres sont calculés par ASCAT à partir de données SNP: L'estimation de la cellularité et l'estimation de quel allèle a été gagné/perdu par rapport à l'autre. Pourront-ils être déterminés à partir de données Oncoscan?)

- ~~regarder si j'ai pas les guidelines pour le rapport quelque part dans moodle. sinon demander à la team geneco, puis à Slim karkar~~
- ~~compléter le document partagé~~

Définition d'un copy number-neutral event dans ASCAT_article.pdf : -> *We define here a copy number-neutral event as an allelic bias for an SNP heterozygous in the germline such that the total copy number does not differ from the tumor ploidy*

arrivée à 09h50; départ à 18h00; pas de pause

Mercredi 23/03/2022

todo: finir de noter pour ASCAT, il reste 2 phrases. faire slides CGHcall. 1 couleur par ppt! Eacon en bleu, OncoscanR en rouge, CGHcall en vert... Objectif aujourd'hui: slides complètes pour CGHcall. 1 slide titre 1 slide pour présenter l'objet cghRaw? revoir de quel package il vient. 1 slide description globale: ce que permet ce package; ses spécificités par rapport aux autres. 1 slide pipeline: comment le fait-il. organigramme avec carrés et flèches 1 slide par étape pertinente du pipeline

je regarde OncoscanR 5 minutes, avec quels documents puis-je en apprendre plus sur ce package? il n'a pas d'article propre, mais cite un article pour chaque score HRD qu'il calcule. voir l'api ou l'aide de R (les 2 contiennent les mêmes infos) pour en savoir plus sur chaque fonction.

arrivée à 10h20

Jeudi 24/03/2022

Objectif aujourd'hui: slides complètes pour EaCon.

Vendredi 25/03/2022

Lundi 28/03/2022

Mardi 29/03/2022

Objectif: avoir les slides pour chaque package. ou à la rigueur 2 bien en profondeur et les 2 autres + tard.