Radiation Desiccation Response Regulated genes in the Radioresistant Bacterium Deinococcus radiodurans

Santiago Holguin Urbano

Analyse de données multi-omiques

Table de matières

Introduction

Bacterium deinococcus radiodurans est une bactérie possédant des caractéristiques génomiques et protéiques leur permettant de survivre dans une condition de radioactivité. Cette résistance est médiée par une réponse rapide de la machinerie transcriptionnelle, qui passe par un changement structurel de la protéine Irre, et par une dégradation de l'inhibiteur DdrO par cette dernière. Suite à cette dégradation, DdrO laisse libre des gènes spécifiques impliqués dans une réponse ciblée aux conditions radioactives. En 2021, une analyse génomique complète de cet organisme a été réalisée par des chercheurs de l'I2BC.

Cette analyse a fourni une liste de gènes candidats comme régulés par le facteur de transcription DdrO. Cette liste a été obtenue en suivant 3 différentes analyses : D'abord une analyse transcriptomique visant à déterminer quels sont les gènes exprimés en absence de DdrO, puis un Chip Seq des séquences que DdrO fixe suivi de l' étude des régions consensus fixés par cette protéine.

Ce compte rendu vise à donner une deuxième approche pouvant complémenter les résultats retrouvés en 2021. Les mêmes données sont utilisées mais des variations dans certaines méthodes de traitement et analyse peuvent varier. Pour s'y faire, nous allons tout d'abord analyser les données de Chip Seq et retrouver les motifs de fixation, afin d'obtenir les séquences ayant été retrouvées en ChipSeq et possédant également un des motifs candidats. Nous allons ensuite poursuivre avec l'analyse du transcriptome tout en croisant les données avec les résultats ChipSeq. Ce même processus dans l'article original résulte en une liste de 37 gènes.

Chip-seq

Peak Calling

ChIP-Seq (Chromatin ImmunoPrecipitation Sequencing) est une technique pour identifier les sites de liaison des protéines sur l'ADN dans le génome. Ce processus se déroule ainsi : l'incubation de la protéine avec l'ADN, formation de ponts stables avec formaldehyde, digestion enzymatique (ou sonication) de l'ADN, immunoprécipitation de la protéine lié à ce ADN, et séquençage des fragments d'ADN co-immuno précipités avec l'anticorps. Afin d' éviter les faux positifs et s'assurer que les pics détectés correspondent bien aux véritables sites de liaison de la protéine d'intérêt, deux contrôles sont réalisés. L'input est l'ADN extrait

avant immunoprecipitation (ChIP), mais traité de la même manière que l'échantillon ChIP (fragmentation, crosslinking, séquençage). Et le mock est l'immunoprécipitation réalisée avec un anticorps contrôle (ex. anticorps contre une protéine absente ou non spécifique).

Le chip-seq a donc résulté en 5 fichiers : IP 1-3, le mock et le input. Pour cette analyse, le IP3 n'est pas utilisé. Le traitement de ces fichiers passe par un premier alignement des fichiers IP avec la séquence de référence, puis par un "Peak calling". Afin de traiter, et puis intégrer les résultats pour les différentes expériences, contrôle et non contrôle, nous avons réalisé deux workflows sur Galaxy.

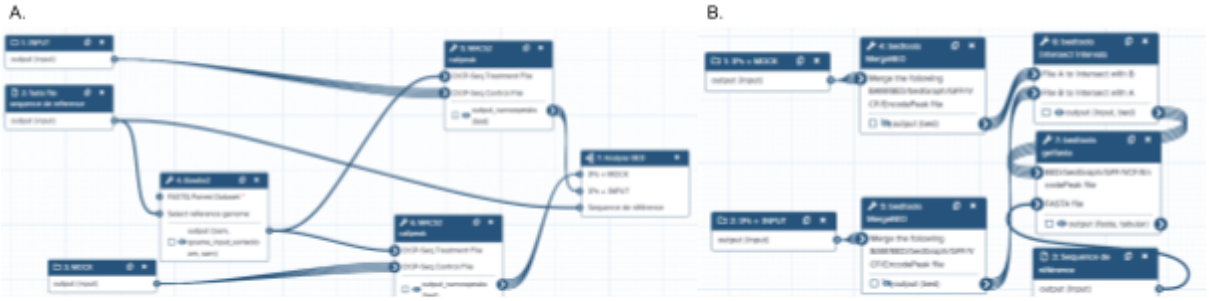


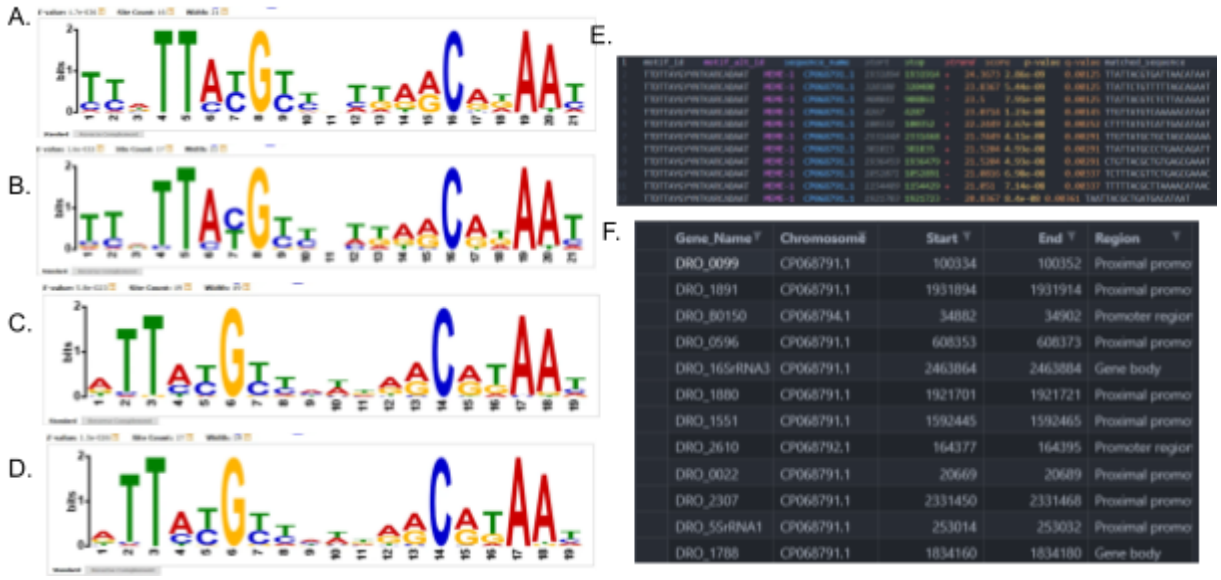
Figure 1 : Workflow d'analyse de données Chip-seq. (A) Workflow galaxy réalisant deux "Peak calling", un pour le MOCK-IP et un pour le INPUT-IP. Ce dernier workflow connect un deuxième (B) traitant ces fichiers. Ce Workflow va réunir en un seul fichier BED tous les Peaks retrouvés par MOCK et séparément ceux par INPUT. Une intersection des résultats est réalisé, puis le fichier bed de cette intersection est utilisée pour obtenir un fichier fasta contenant les séquences d'ADN.

Le premier workflow aligne IP et séquence de référence, et trouve ensuite les peaks pour l'input et pour le mock (Figure 1.A). Le deuxième "merge" tous les peaks trouvés avec chaque type de contrôle, puis fait une intersection entre les deux contrôles. Cela nous donne comme résultat un fichier bed et un fichier fasta, avec 452 séquences. Les auteurs de l'article on réalisé le peak calling en utilisant une alternative à MACS2, ils ont en effet utilisé B-peaks, une librairie R. Ainsi, ils ont retrouvé 136 séquences.

Recherche de motifs

La recherche de motifs peut se faire par différents algorithmes. MEME est un des nombreux algorithmes qui peut être utilisé, mais possède plusieurs limites. En effet, entre plus de séquences à analyser il y a, plus de temps à retrouver les séquences l'algorithme va prendre, et moins de fiabilité. Une alternative est donc de faire avant une recherche de motifs à partir de motifs de référence, en utilisant l'algorithme FIMO. Ces dernières séquences de référence correspondront donc aux séquences de gènes ayant été démontrés expérimentalement comme étant fixés par DdrO.

Nous avons donc d'abord obtenu 4 matrices de référence en utilisant ces 16 séquences. Ces 4 matrices correspondent à la variation de deux paramètres : à 0 ou 1 occurrence par séquence (Figure 2.A,D) ou n'importe quelle nombre d'occurrence (Figure 2.B,C), et palindromique (Figure 2.C,D) ou non palindromique (Figure 2.A,B). Ces motifs sont ensuite recherchés dans nos 452 séquences, donnant ainsi comme résultat 4 fichiers tsv avec 10 colonnes (figure 2.E).



Gene_Name	Chromosome	Start	End	Region
DRO_0099	CP068791.1	100334	100352	Proximal promo
DRO_1891	CP068791.1	1931894	1931914	Proximal promo
DRO_B0150	CP068794.1	34882	34902	Promoter regior
DRO_0596	CP068791.1	608353	608373	Proximal promo
DRO_16SrRNA3	CP068791.1	2463864	2463884	Gene body
DRO_1880	CP068791.1	1921701	1921721	Proximal promo
DRO_1551	CP068791.1	1592445	1592465	Proximal promo
DRO_2610	CP068792.1	164377	164395	Promoter regior
DRO_0022	CP068791.1	20669	20689	Proximal promo
DRO_2307	CP068791.1	2331450	2331468	Proximal promo
DRO_5SrRNA1	CP068791.1	253014	253032	Proximal promo
DRO_1788	CP068791.1	1834160	1834180	Gene body

Figure 2 : Résultats de la recherche de motifs. L'algorithme FIMO nous donne comme résultat 4 matrices en fonction des paramètres utilisés : 01 et non palindromique (A), any-number et non palindromique (B), any-number et palandromique (C), 01 et palindromique (D). Le recherche de motifs dans nos données de chip seq résulte dans un fichier tsv (E), qui est ensuite utilisé pour retrouver les gènes fixés et contenant un des motifs (F).

Afin de retrouver à quels gènes sont associées chacune de ces séquences, un pipeline d'analyse est réalisé en utilisant un script Python (LIEN GITHUB). Ce script va donc retrouver les gènes ayant le motif recherché en utilisant les paramètres suivants : Le motif se retrouve dans le partie intergénique du promoteur (entre la fin du gène et le début du gène d'intérêt), dans le corps du gène (soit entre lé début et la fin si la longueur du gène est supérieure à 100, soit entre le début et 100 nucléotides après si non), ou dans la région promotrice proximale (-500 nt avant le TSS). Nous obtenons donc un fichier dans le format présenté dans la figure 2.F, contenant 288 gènes candidats sur 452 gènes initiaux.

Cette valeur est largement supérieure aux 110/136 gènes retrouvés pendant l'étude. De ces 288 gènes 19 (comme dans l'article) sont des gènes de référence (sur 25). Si on fait le même filtrage sur les 136 gènes que sur nos 452, nous obtenons non pas 110 gènes mais 84 gènes qui se retrouvent dans une région promotrice. Cette variabilité peut s'expliquer par notre choix de dire que si le gène a une longueur x de <100 nt, les fixation après ces x nucléotides ne sont pas comptés comme des possibles sites de régulation; mais aussi par des variabilités dans les analyses PeakCalling.

Ces choix sont justifiés pour augmenter la précision par rapport au niveau de fixation de DdrO, et pour augmenter la flexibilité par rapport à ce qui peut ou ne peut pas être une immunoprécipitation correcte par rapport au contrôl, sans sortir des standards (pv < 0.05 sur MACS2) mias juste en changeant la méthode.

Transcriptome

L'étude des gènes différentiellement exprimés en conditions d'absence de DdrO passe par une analyse du transcriptome en comparant les différentes conditions sans DdrO, avec la condition avec DdrO. Cette première condition correspond à xH après le changement de température restreignant la réplication du plasmide qui contient la seule séquence codant pour une protéine Ddro dans la cellule (Figure 3.A) . La deuxième correspond au transcriptome extrait 1H après l'augmentation de la température (Figure 3.B).

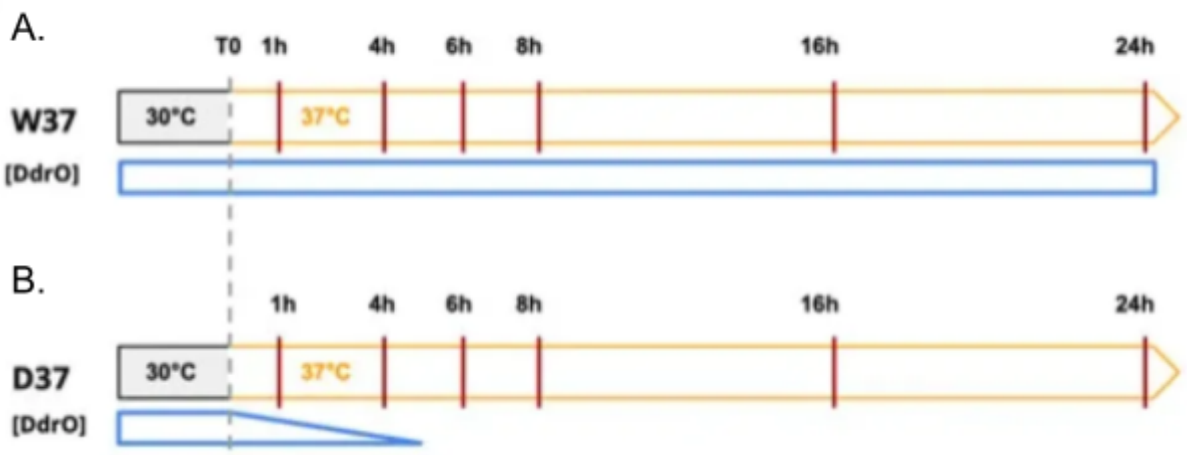


Figure 3 : Procédure pour l'extrait d'ARN cellulaire. Évolution de la température et de la quantité estimée de DdrO pour la souche W37, contenant un plasmide non thermosensible (A), et pour la souche D37 contenant un plasmide thermosensible (B). Les deux souches sont KO pour DdrO. Figure issue de l'article (Eugénie et al. 2021).

Les données transcriptomiques suivent un pipeline qui est transformé en workflow Galaxy (Figure 4). Ce pipeline nous permet de faire un contrôle qualité avec FASTQC, un alignement avec bowtie, un comptage de reads par gène par htseq-count, et une analyse différentielle avec DESEQ2 entre les conditions 1H contre Xh. Toutes les analyses sont faites ainsi, nous donnant comme résultat 10 fichiers tsv qui sont transformés en un seule utilisant le script “extrait_donnees.R” (LIEN GITHUB).

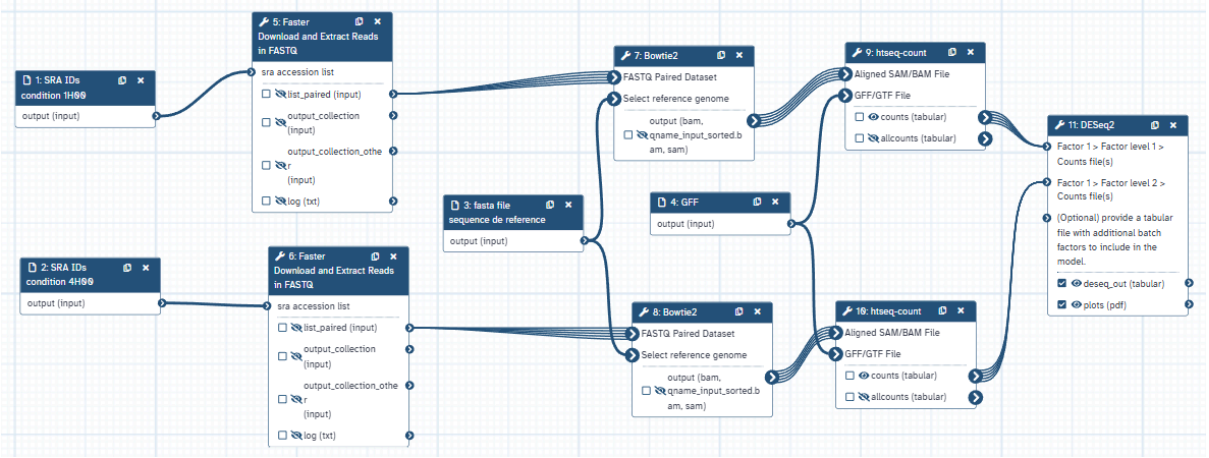


Figure 4 : Workflow d’analyse de transcriptome sur Galaxy et traitement.

Analyse double

Préparation de données et paramètres choisies

Contrairement à l'article, où les résultats sont retrouvés et filtrés séparément avant d'être intersectés ou fusionnés, nous avons opté pour une analyse double. Dans cette approche, le paramètre transcriptome influence le filtrage du ChIP-seq et vice-versa. Dans les sections précédentes, nous avons d'abord augmenté notre jeu de séquences fixées afin de ne pas perdre des séquences potentiellement peu fixées par DdrO. Ensuite, nous avons traité les données transcriptomiques en calculant le log2FC et la p-value (ajustée et non ajustée). Plutôt que de filtrer ces résultats séparément, nous avons adopté un filtre global.

Pour préparer les données, nous avons créé un dataframe regroupant plusieurs paramètres essentiels au filtrage (Tableau 1). Cela permet de réduire à une seule ligne par gène les informations clés pour déterminer si un gène est fixé par DdrO (présence du motif et détection en ChIP-seq) et s'il est surexprimé en absence de DdrO. Contrairement à l'article, nous avons analysé la p-value ajustée pour quantifier combien de fois elle était significative

en condition de surexpression en l'absence de DdrO, mettant ainsi en avant les gènes qui varient positivement par rapport au contrôle.

Colonne	Description
FC_sup_2.5_D37	Nombre de fois où le `log2FC` est supérieur à 2,5 pour l'échantillon "D37"
FC_sup_1.5_D37	Nombre de fois où le `log2FC` est supérieur à 1,5 pour l'échantillon "D37"
FC_sup_2_4fois_D37	Indique si le `log2FC` est supérieur à 2 au moins 4 fois pour l'échantillon "D37".
FC_inf_1.5_W37	Nombre de fois où le `log2FC` est inférieur à 1,5 pour l'échantillon "W37".
FC_inf_1_4fois_W37	Indique si le `log2FC` est inférieur à 1 au moins 4 fois pour l'échantillon "W37" (valeur booléenne).
Significatif_diff	Différence entre le nombre de fois où le `log2FC` est significatif (`P_adj < 0,01`) et supérieur à 1 pour "D37" par rapport à "W37".
Associated	Indique si le gène fait partie des gènes issues du Chip Seq et contenant le motif
Reference	Indique si le gène est dans la liste des gènes de référence.
Associated_region	Région de fixation de DdrO sur le gène en fonction du motif (Proximal promoter, promoter region, Gene body, ou NOT).

Tableau 1 : Paramètres choisis pour le filtrage des gènes.

Filtrage (script filtrage.R github)

Pour filtrer les gènes selon des conditions spécifiques, nous avons créé une nouvelle colonne "Condition" dans notre dataframe. Cette colonne attribue une valeur numérique à chaque gène en fonction de critères prédéfinis, permettant ainsi de catégoriser les gènes

selon leur comportement d'expression et leur pertinence biologique. Si le gène ne suit aucune condition un valeur de -2 est donné dans conditions (moyenne négative des 4 conditions importante pour l'ACP qui est réalisé en amont).

La première condition, 51 gènes, regroupe donc les gènes qui se sont fortement exprimés en absence de DdrO pendant un temps court et dont leur expression en présence de DdrO ne peut augmenter que très très peu dans une seule condition maximum (’entre 1 et 1.5, soit 50% de ce qui a varié en absence de DdrO). La condition 2, 16 gènes, englobe les gènes ayant varié pendant plus longtemps mais peut-être d'une manière moindre, tout en gardant une faible augmentation dans les conditions de contrôle. La condition 3, 26 gènes, comprend les gènes ayant le motif et ayant été présents dans le Chip seq, mais ayant eu au moins un peu plus de transcrits en absence de DdrO : seuls 3 gènes. La condition 4, 2 gènes, englobe les gènes où l'expression a augmenté faiblement pendant longtemps tout en n'ayant presque pas du tout varié dans les conditions contrôles.

Condition	Description
1	`log2FC` > 2,5 au moins une fois pour "D37", `log2FC` < 1 au moins quatre fois pour "W37", `log2FC` < 1,5 tout le temps pour "W37", et différence significative d'au moins trois fois.
2	`log2FC` > 1,5 au moins trois fois et > 2,5 au moins deux fois pour "D37", `log2FC` < 1,5 au moins trois fois pour "W37", et différence significative d'au moins trois fois.
3	`log2FC` > 1,5 au moins une fois pour "D37", `log2FC` < 1 au moins quatre fois pour "W37", gène associé, et différence significative d'au moins une fois.
4	`log2FC` > 2 au moins quatre fois pour "D37", `log2FC` < 1 au moins quatre fois pour "W37", et différence significative d'au moins trois fois.

Tableau 2 : Différents types de filtres rendant un gène comme candidat.

Ce filtrage nous donne ainsi un résultat de 96 gènes candidats, contre 35 dans l'article. Avec seulement 10 gènes se trouvant dans les 2 résultats. Ces résultats peuvent paraître moins bons parce que peu de ressemblance existent, c'est donc la raison pour laquelle nous avons décidé de faire une comparaison entre ces résultats en utilisant comme point de comparaison les caractéristiques des séquences de référence.

Comparaison des résultats et discussion

Afin de comparer les résultats de l'article avec les nôtres, nous avons décidé de réaliser une analyse en composantes principales. Bien que cette analyse puisse être biaisée parce que les variables expliquant le modèle sont les paramètres de filtrage de nos gènes, nous avons justifié le choix de ces variables dans un contexte biologique global et devraient pouvoir aussi s'appliquer aux gènes retrouvés dans l'article. Deux analyses sont réalisées, la première prend en compte toutes les variables numériques et même si on pourrait considérer comme "équilibré", le biais qu'elle pourrait induire ne nous permet pas de l'utiliser pour comparer mathématiquement les résultats mais seul pour avoir une idée graphique des résultats. La deuxième ACP enlève tous les biais en calculant les coordonnées uniquement à partir des variables globales, sans les variables propres à chaque groupe (on enlève par exemple la variable Condition ou la variable Article).

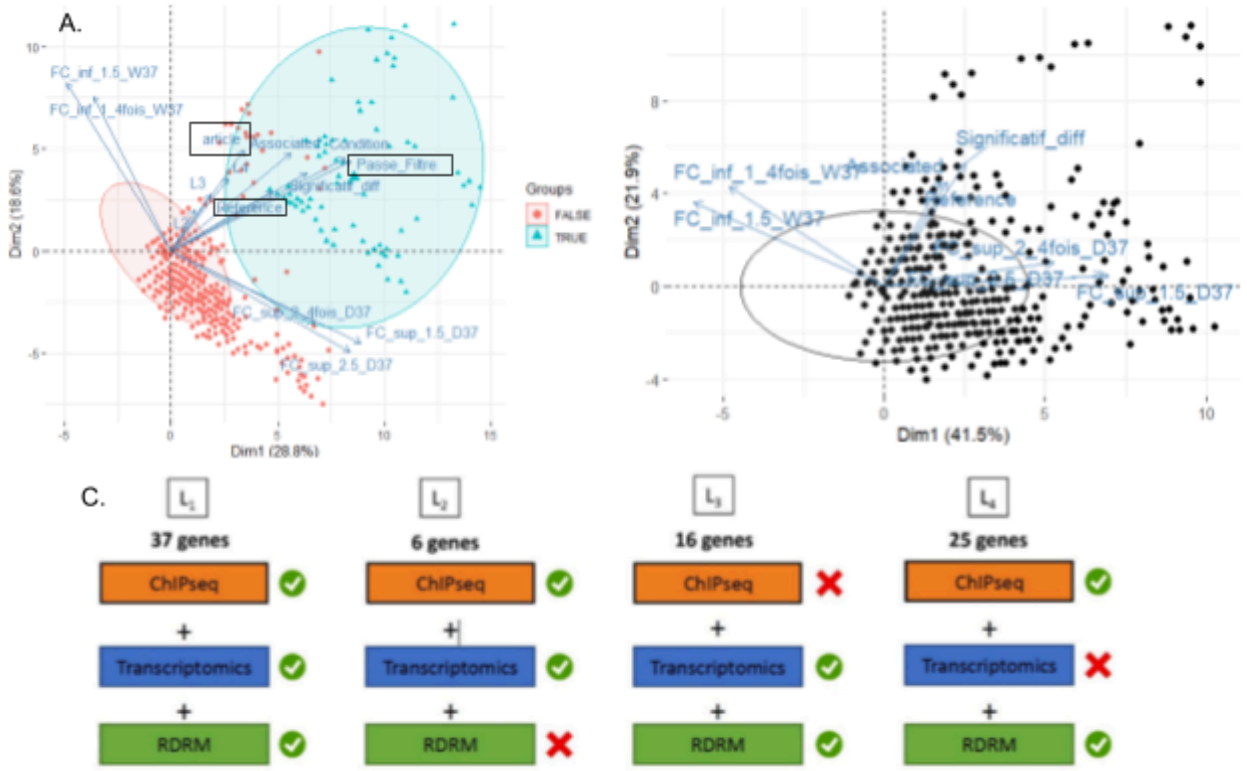


Figure 5 : Analyses en composantes principales sur l'ensemble de données. La première (A) est réalisée avec l'ensemble de variables numériques : toutes celles en Tableau 1 sauf Associated_region, plus L2, L3, L4, article, condition, Passe_filtre. L2-4 correspondent aux groupes dans l'article (C) et article est une variable booléenne. Le deuxième analyse (B) est uniquement réalisé avec les variables globales : toutes celles en Tableau 1 sauf Associated_region.

En figure 5.A nous observons que les gènes issues de l'Article et les nôtres (Passent_filtre) suivent bien la même tendance que ceux de référence. Graphiquement, nous pouvons constater que nos gènes sont plus proches du vecteur représentant les gènes de référence. Nos gènes semblent paradoxalement englober les gènes fortement exprimés tandis que les gènes de l'article deux dont la variation est moindre dans les conditions contrôle et plus dans les séquences Chip Seq. Le paramètre Signifactif_diff semble être le paramètre expliquant le mieux les gènes référence.

Pour l'ACP sans biais (ou avec moins), graphiquement rien ne peut être commenté. Cependant, on a utilisé le résultat de cet ACP pour calculer la distance euclidienne entre reference-article et reference-Passsent_filtre (nos gènes). Ce calcul résulte dans une distance de 12.6 et 11.8 respectivement, nous indiquant que dans un espace vectorielle définie par des variables explicatives telles que celles montrées en Tableau 1, les gènes que nous avons trouvés suivent une tendance plus ressemblante aux gènes de référence que les gènes retrouvés dans l'article. La liste des gènes filtrés peut-être retrouvée sur Github dans le fichier "resultats_filtres.csv".

Afin d'étudier si l'algorithme de PeakCalling pourrait modifier quelque chose, nous avons réalisé le même script de filtrage mais à partir de ce fichier. Nous trouvons cette fois-ci 71 gènes, et un résultat similaire: les distances euclidiennes sont cette fois-ci 11.6 pour nos gènes et 12 pour les résultats de l'article. Ce changement réduit la distance globale et rapproche les deux résultats. Cela pourrait s'expliquer parce que l'algorithme b peaks est spécifique pour les procaryotes tandis que MACS2 est plus spécifique des eucaryotes, le rendant donc moins adapté pour cette analyse.