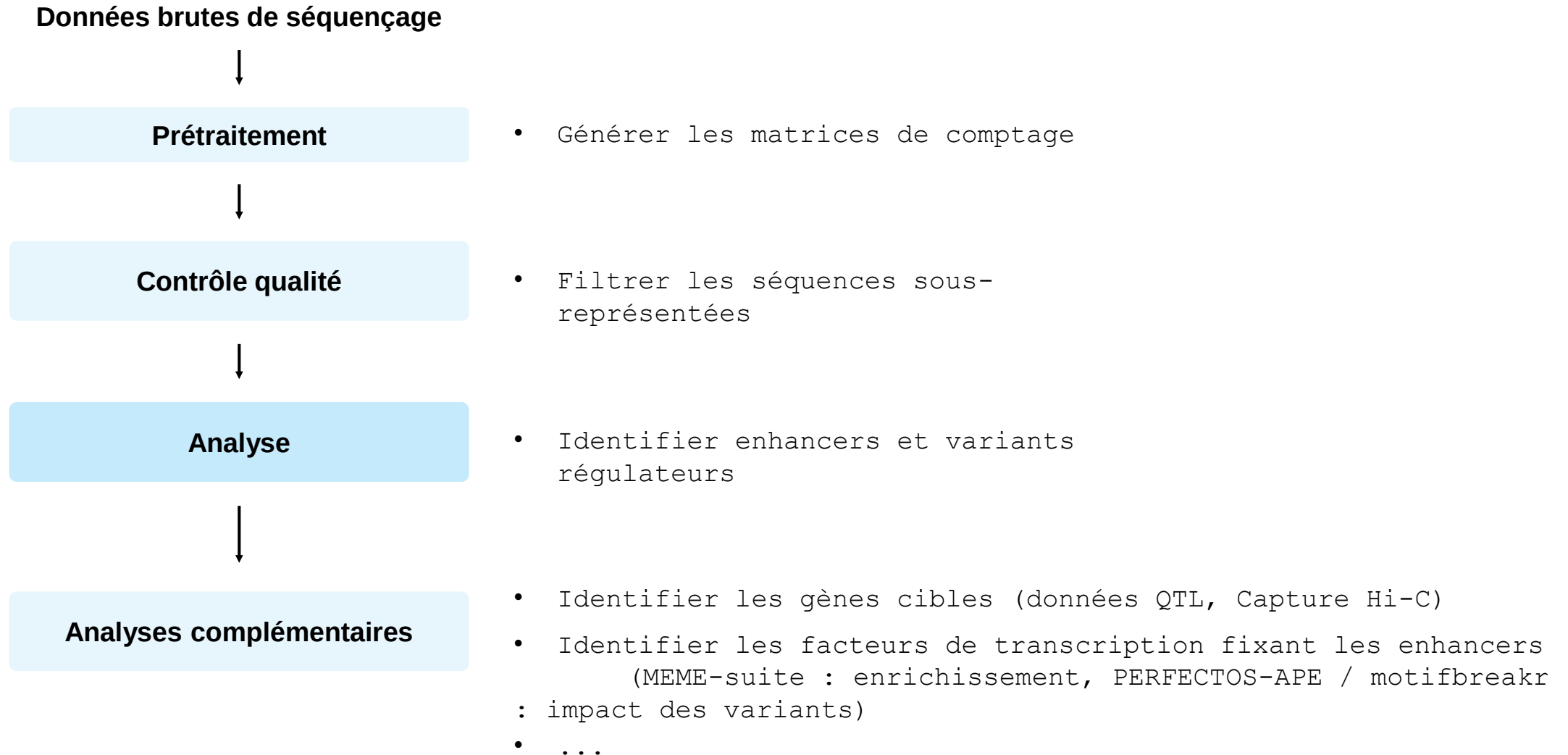
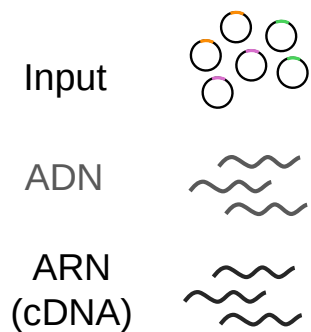


Analyses des données STARR-seq



Prétraitement des données*



Fichier FASTQ



Filtre reads <140nt
Cutadapt



Alignement
Bowtie2



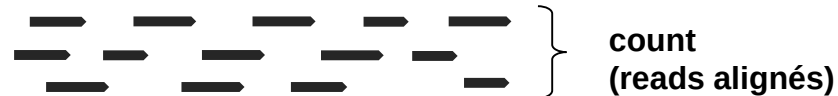
Filtre reads non alignés et reverse
+ tri et indexation
SAMtools

93–
99%



Quantification
du nombre de reads par séquence

Séquence testée



Matrice de comptage (count matrix)

	Replicat 1	Replicat 2	Replicat 3	Replicat 4
Sequence 1	1339	1391	1152	1451
Sequence 2	1693	1607	1588	1580
Sequence n	...	...	...	...
Sequence 12000	440	454	446	338

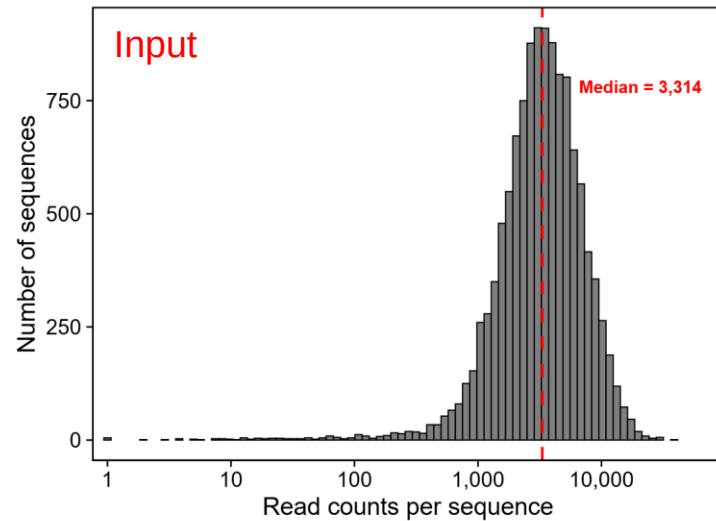
2 matrices : une pour l'ADN et une pour l'ARN

*Cette étape a été réalisée par Nori Sadouni

Contrôle qualité

➤ Librairie correctement représentée ?

Distribution des counts

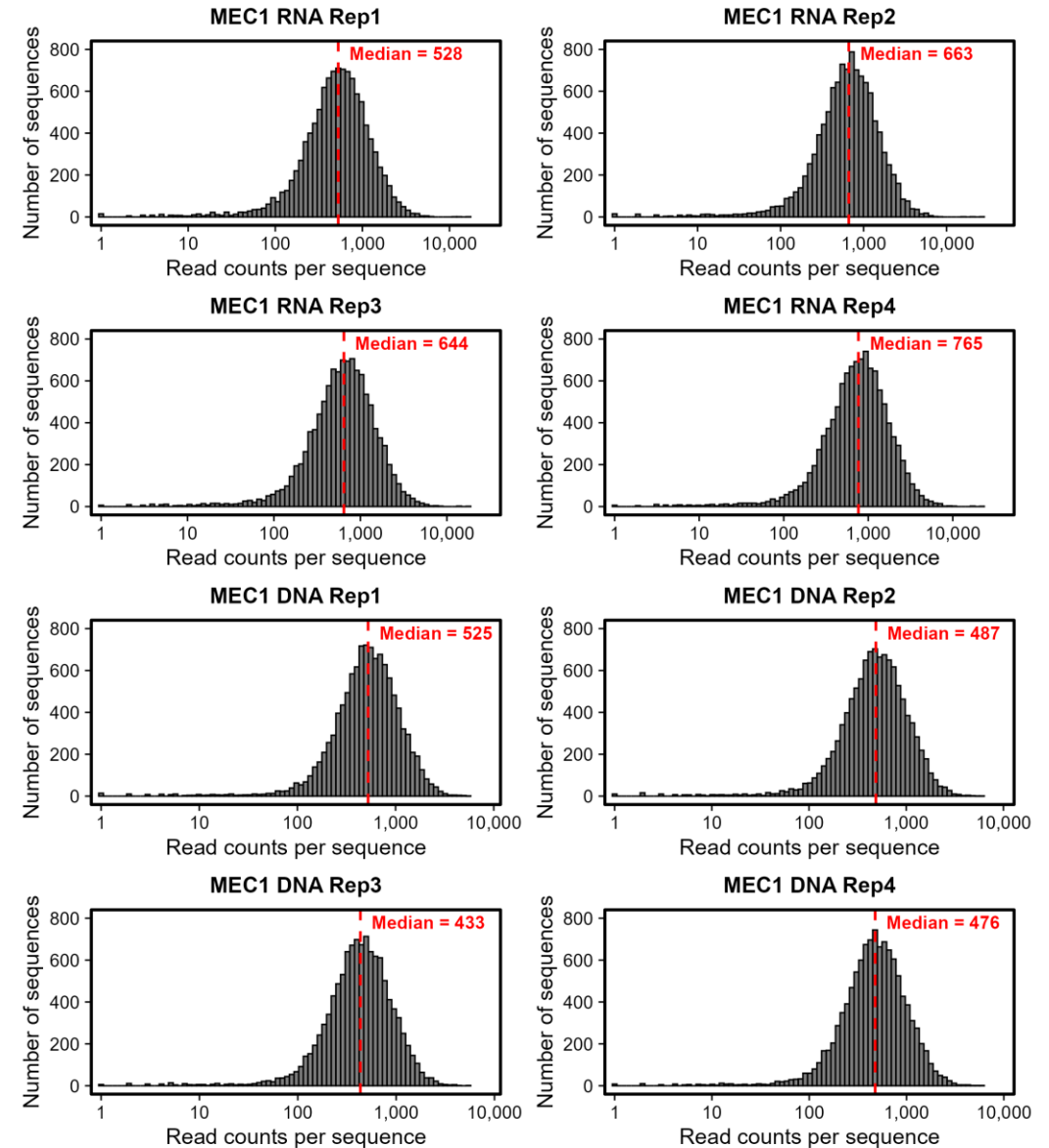


Séquence non représentée : count input ou ADN= 0

Séquence peu représentée : count très faible

➤ Filtre sur les counts ?

Réplicats



Contrôle qualité

➤ Filtre sur les counts ?

Séquence non représentée : count input ou ADN = 0

Séquence peu représentée : count très faible

exemple :

	ADN Rep1	ADN Rep2	ADN Rep3	ADN Rep4	Input
Sequence 1	20	19	4	15	38
Sequence 2	1	1	0	0	5
Sequence 3	8	5	5	9	25
Sequence 4	0	0	0	0	0

Quelles séquences faut-il garder ?

Contrôle qualité

➤ Filtre sur les counts ?

Exemple de seuils choisis dans diverses publications :

Myint et al. 2019 (**ADN < 10**)

Sun et al. 2023 (**Barcode < 5**)

Hammar et al. 2025 (**CPM < 1**)

Zhang et al. (**ADN < 10**)

Pas de précision :

« barcodes /fragments with low representation were removed prior to activity analysis »

Zhang et al. *Genome Biol.* 2025 : comment la variance évolue par rapport à la moyenne des counts pour chaque seuil ? (STARR-seq sur génome entier : librairie complexe)

Contrôle qualité

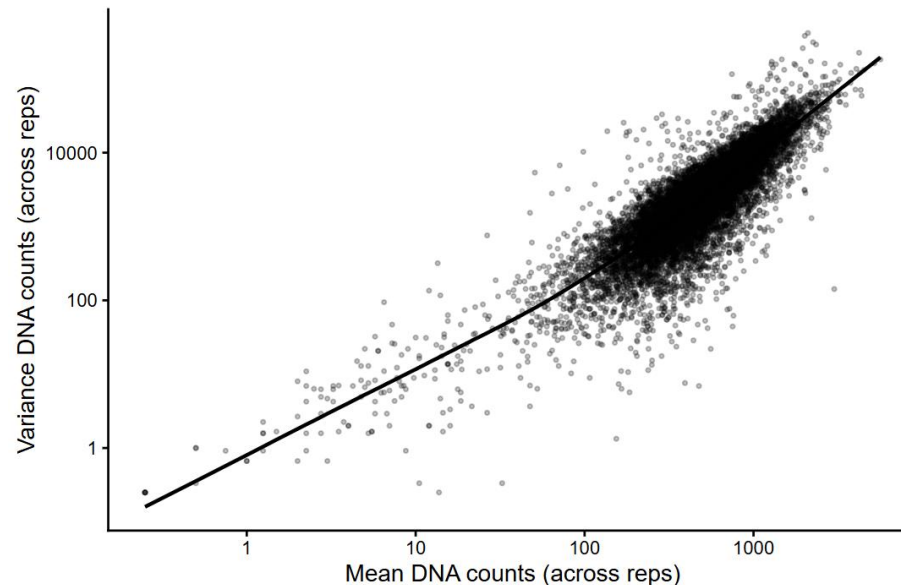
➤ Filtre sur les counts ?

Zhang et al. *Genome Biol.* 2025 : : comment la variance évolue par rapport à la moyenne des counts pour chaque seuil ? (recommandation pour du STARR-seq sur génome entier : librairie complexe)

x = moyenne des counts d'ADN

y = variance des counts d'ADN

Même si notre librairie STARR-seq n'est pas complexe (~12 000 séquences), on peut suivre les recommandation de Zhang et al. :



La relation entre la moyenne et la variance se rapproche d'une distribution binomiale négative ce qui est attendu

Contrôle qualité

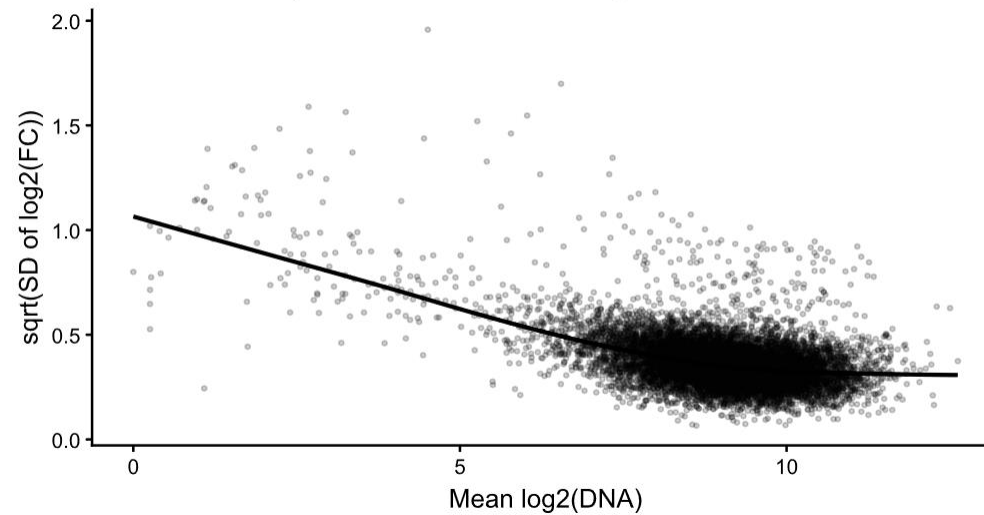
➤ Filtre sur les counts ?

La variabilité entre réplicats diminue lorsque les counts d'ADN augmentent comme le montre la courbe de regression.

Regression LOWESS

x = moyenne des counts d'ADN d'une séquence

y = variabilité de l'activité de la séquence ($\log_2(FC)$)



On peut donc chercher le seuil qui permet de diminuer la variabilité entre les réplicats en évaluant la corrélation de l'activité des séquences entre réplicats et son écart-type (SD) ...

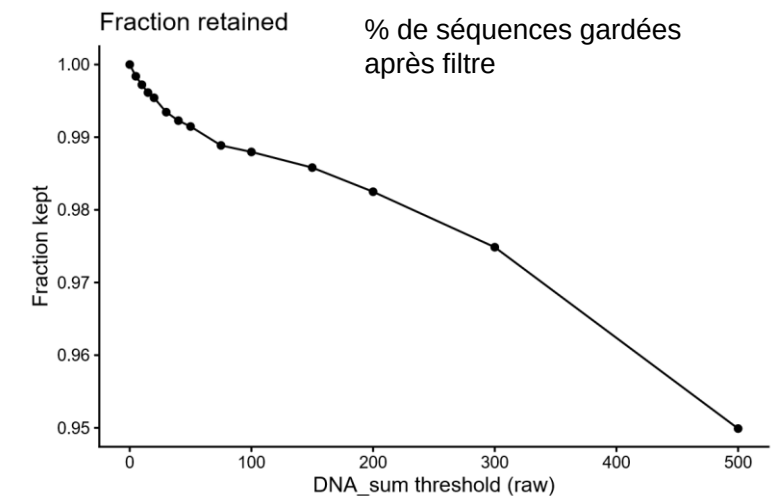
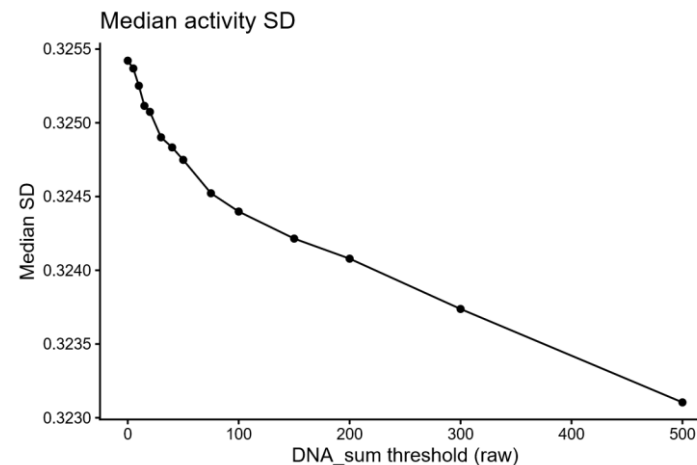
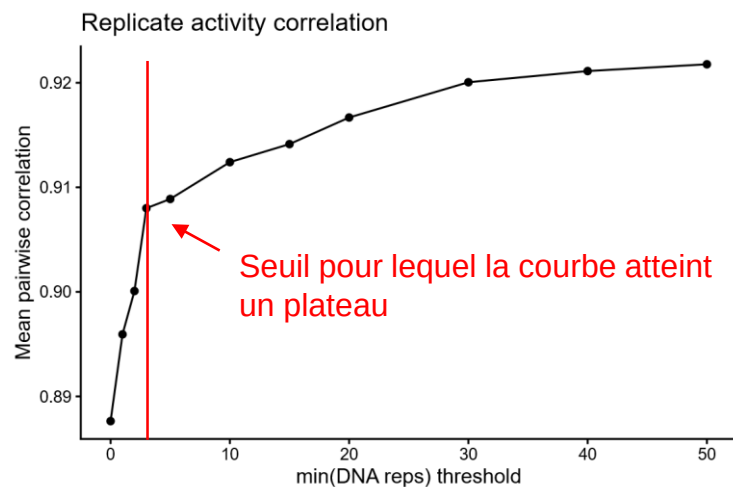
Contrôle qualité

➤ Filtre sur les counts ?

... on cherche donc à améliorer la corrélation entre les réplicats et diminuer l'écart-type (SD) tout en conservant le maximum de séquences dans la librairie

1. **Calcul de l'activité de chaque séquence :** $\log_2(FC) = \log_2 \frac{ARN}{ADN}$
2. **Corrélation de l'activité entre réplicats :** coefficient de Pearson R
3. **count ADN > seuil** pour tout les réplicats

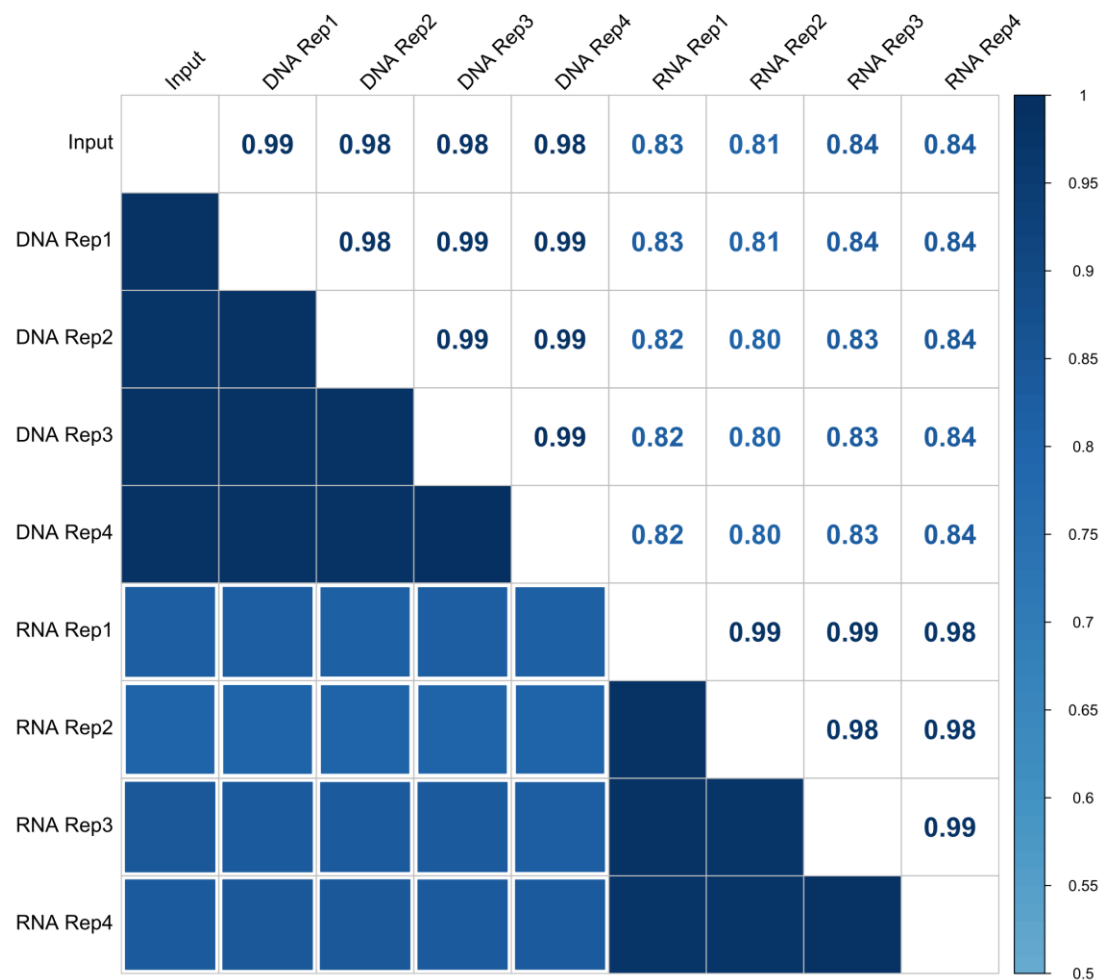
Pour nos données, le coefficient de corrélation R est déjà très élevé car on a très peu de séquences avec de faibles counts. Fixer un seuil n'impacte pas énormément le coefficient.



Contrôle qualité

➤ Expérience reproductible ?

Corrélation (R) des counts entre réplicats



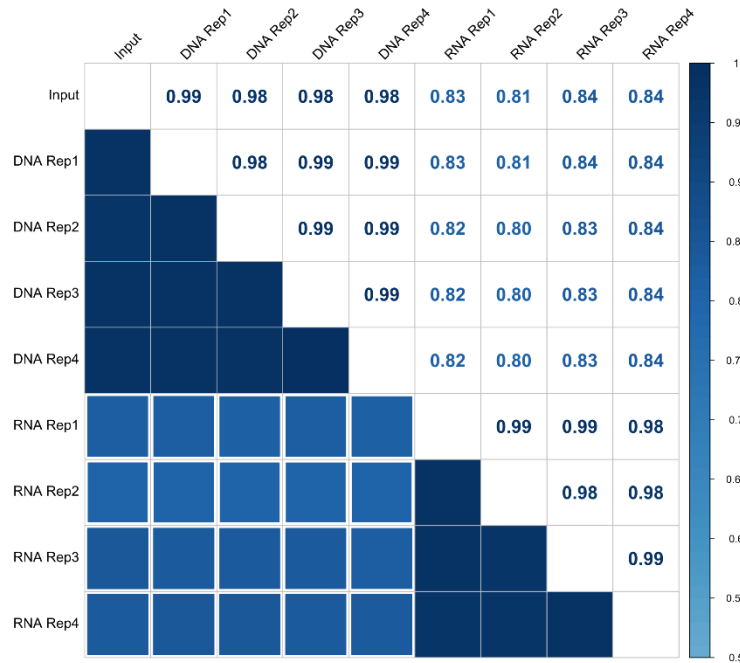
Possibilité d'utiliser :

- Counts non normalisés
- Normalisation (CPM, RPM...)
- Transformation $\log_{10}(\text{counts}) + 1$, $\log(\text{CPM})$...

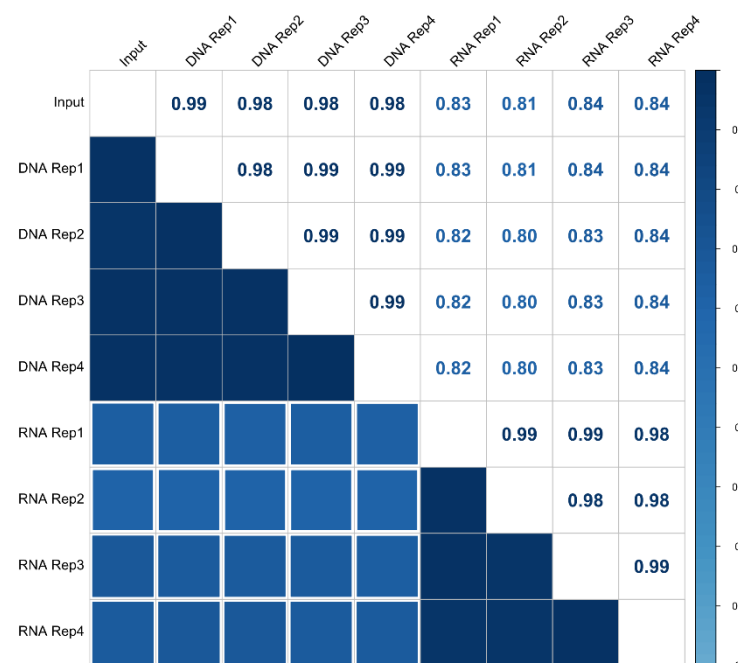
Contrôle qualité

➤ Expérience reproductible ?

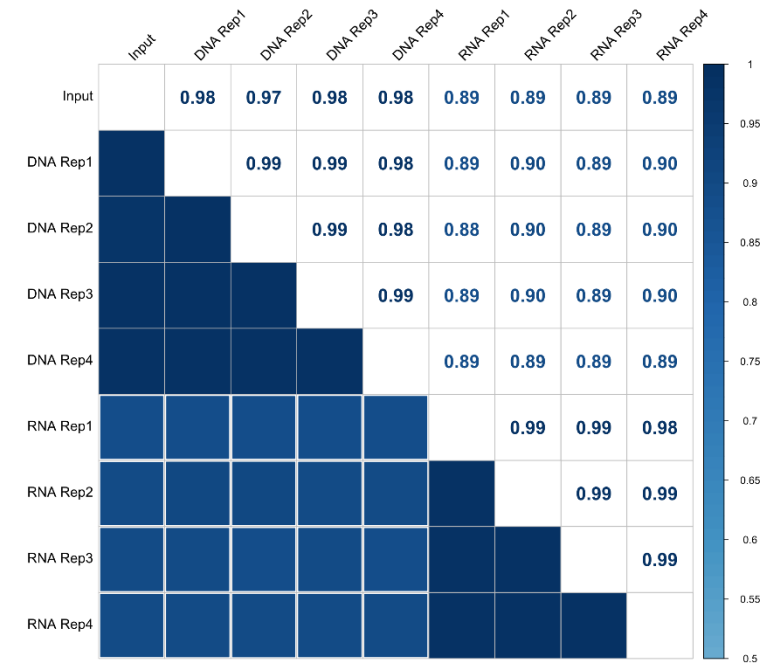
Pas de grandes différences entre counts normalisés ou non



- Counts brutes



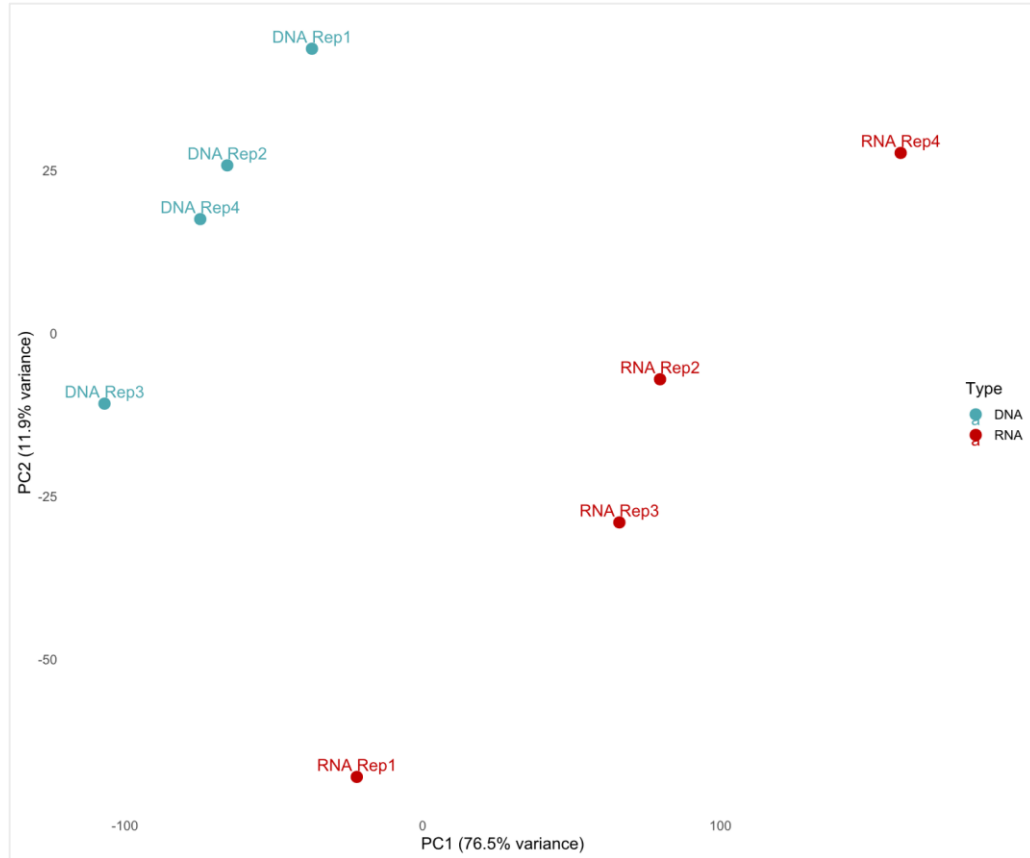
- CPM



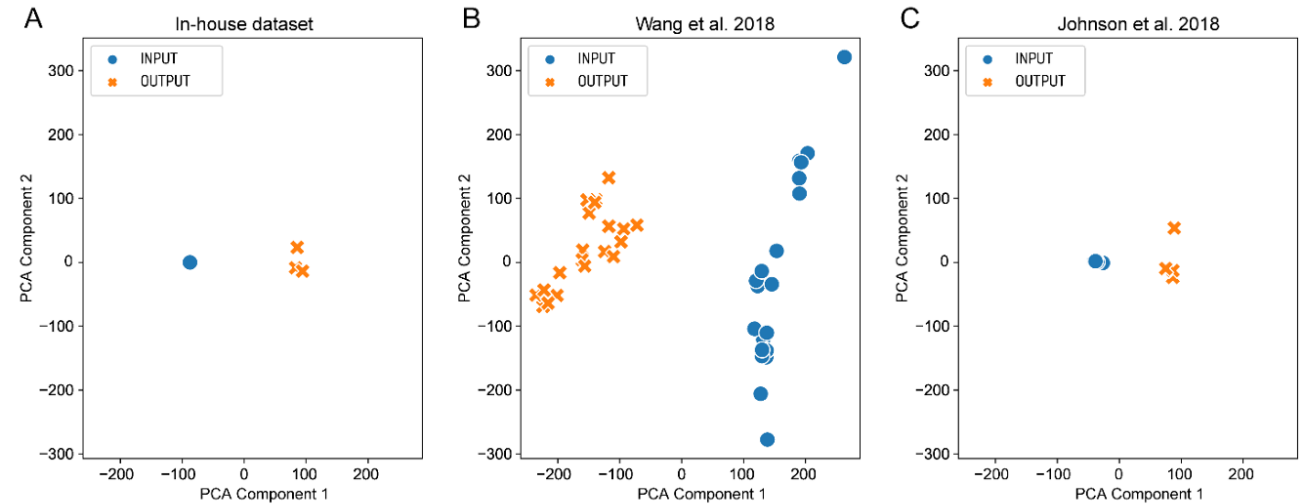
- Log10(counts) + 1

Contrôle qualité

Analyse en composantes principales (ACP)



Supplemental Fig. S5 (A-C): Principal Component Analysis (PCA) for STARR-seq library replicates

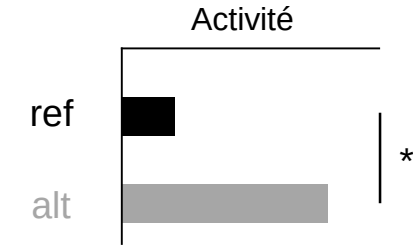
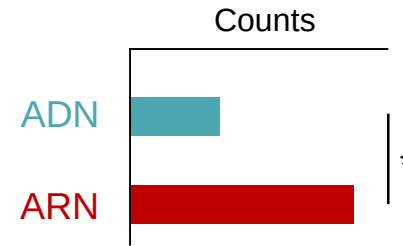


Das et al. *Challenges and considerations for reproducibility of STARR-seq assays*. Genome Res. 2023,

Analyse

Objectifs :

- Identifier les séquences enhancers
- Identifier les variants régulateurs



➤ Quelles méthodes utiliser ?

Simple ratio : $\log_2(\text{FC}) = \log_2 \frac{\text{counts ARN}}{\text{counts ADN}}$

RNA-seq : DESeq2, edgeR, limma

ChIP-seq : MACS2

MPRA : MPRAalyze, mpralm, QuASAR-MPRA...

STARR-seq : STARRPeaker, CRADLE...

Analyse avec MPRAnalyze

Ashuach et al. *Genome Biology* (2019) 20:183
<https://doi.org/10.1186/s13059-019-1787-z>

Genome Biology

METHOD

Open Access

MPRAnalyze: statistical framework for massively parallel reporter assays



Tal Ashuach^{1,2†}, David S. Fischer^{3,4†}, Anat Kreimer^{1,5,6}, Nadav Ahituv^{5,6}, Fabian J. Theis³ and Nir Yosef^{1,2,7,8*}



Home > Bioconductor 3.22 > Software Packages > **MPRAnalyze**

MPRAnalyze

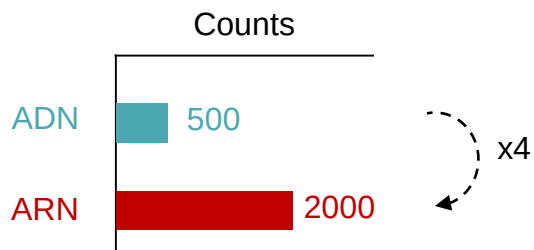
Deux analyses possibles :

1. **Analyse quantitative** : identifier les enhanceurs actifs
2. **Analyse comparative** : comparer deux conditions

Activité d'une séquence

➤ Filtre sur les counts ?

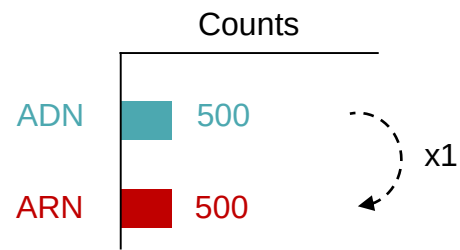
- Calcul de l'activité d'une séquence : $activité = FC = \frac{ARN}{ADN} \longleftrightarrow \log_2(activité) = \log_2 \frac{ARN}{ADN}$



$$activité = \frac{2000}{500} = 4$$

$$\log_2(activité) = +2$$

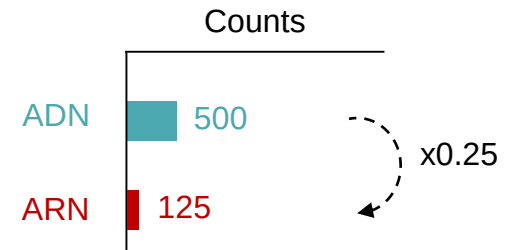
4x plus d'ARN que d'ADN
Séquence enhancer



$$activité = \frac{500}{500} = 1$$

$$\log_2(activité) = 0$$

Autant d'ARN que d'ADN
Séquence inactive



$$activité = \frac{125}{500} = 0.25$$

$$\log_2(activité) = -2$$

4x moins d'ARN que d'ADN
Séquence silencer

Analyse avec MPRAnalyze

Ashuach et al. *Genome Biology* (2019) 20:183
<https://doi.org/10.1186/s13059-019-1787-z>

Genome Biology

METHOD

Open Access

MPRAnalyze: statistical framework for massively parallel reporter assays



Tal Ashuach^{1,2†}, David S. Fischer^{3,4†}, Anat Kreimer^{1,5,6}, Nadav Ahituv^{5,6}, Fabian J. Theis³ and Nir Yosef^{1,2,7,8*}



Home > Bioconductor 3.22 > Software Packages > MPRAnalyze

MPRAnalyze

Deux analyses possibles :

1. **Analyse quantitative** : identifier les enhanceurs actifs
2. **Analyse comparative** : comparer deux conditions

Activité ou taux de transcription

$$\vec{r} \sim NB \left(\mu = \frac{\alpha \cdot k}{\beta}, \psi = k \right)$$

Prend en compte :

- la variabilité entre réplicats
- la profondeur de séquençage
- les contrôles négatifs
- les barcodes (MPRA)
- ...

Détecte uniquement les enhanceurs et séquences inactives (pas les silenciers)

Analyse avec MPRAnalyze

1. Analyse quantitative : identifier les enhanceurs actifs

Fonction : `analyzeQuantification()`

Test statistique: `testEmpirical()`

$H_0 : \alpha_i = \alpha_{\text{neg}} \longleftrightarrow H_0 : \alpha_i = 0$ la séquence i n'a pas d'activité enhancer

Résultats

element_id	statistic	zscore	mad.score	pval.mad	pval.zscore	alpha	FDR	call
rs10005419_ref_C	0,8717	-0,3829	-0,5409	0,7057	0,6491	0,8717	0,9384	False
rs10005419_alt_A	1,0250	0,1796	0,2537	0,3999	0,4287	1,0250	0,8185	False
rs9271166_alt_T	7,6668	24,5370	34,6618	1,48E-263	2,98E-133	7,6668	5,45E-260	True
rs9271165_alt_G	7,2992	23,1890	32,7576	1,18E-235	2,94E-119	7,2992	3,27E-232	True

Pas de correction de tests multiples intégrés

➤ Quel seuil de significativité ?

pval.mad ou FDR

Seuil choisi dans notre étude : $\text{FDR} \leq 0.05$

Analyse avec MPRAnalyze

2. **Analyse comparative** : comparer l'activité des sequences entre deux conditions (avec ou sans stimulation, ref vs. alt...)

Fonction : `analyzeComparative()`

Test statistique: `testLrt()`

$H_0 : \alpha_{i,1} = \alpha_{i,2}$ l'activité de la séquence i ne diffère pas entre la condition 1 et la condition 2 (ex : $H_0 : \alpha_{\text{ref}} = \alpha_{\text{alt}}$)

Résultats

variant	statistic	pval	fdr	df.test	df.dna	df.rna.full	df.rna.red	logFC	FC	Effect	call
rs10005419	7,8218	0,0052	0,0077	1	5	6	5	0,1575	1,1154	1.12× stronger	False
rs1009726	27,2561	1,78E-07	1,20E-05	1	5	6	5	-0,3476	0,7859	1.27× weaker	False
rs2295614	69,6640	7,03E-17	1,23E-13	1	5	6	5	4,5068	22,7336	22.73× stronger	True
rs73258637	14,7244	0,0001	0,0013	1	5	6	5	-4,7689	0,0367	27.26× weaker	True

➤ Quel seuil de significativité ?

FDR et logFC

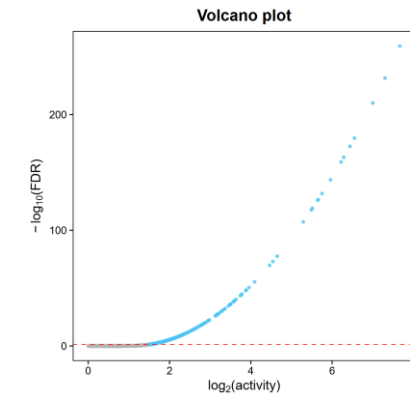
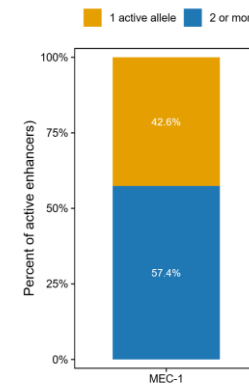
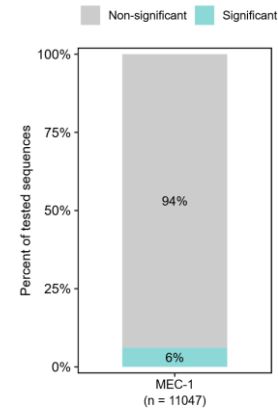
Seuil choisi dans notre étude : $\text{FDR} \leq 0.05$ et $|\log\text{FC}| \geq 0,585$ (1,5x)

Analyse avec MPRAnalyze

1^{ère} méthode

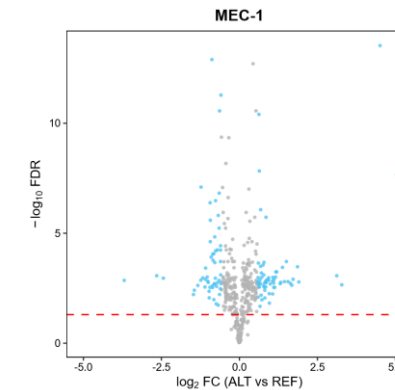
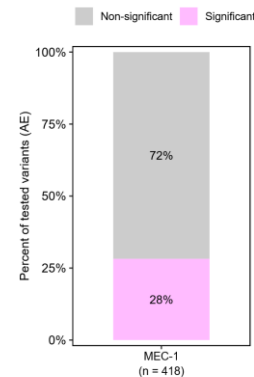
Identification des enhancers

→ 678 / 11 047 séquences avec un effet enhancer



Identification des variants régulateurs dans des enhancers

→ 118 / 419 variants avec un effet régulateur

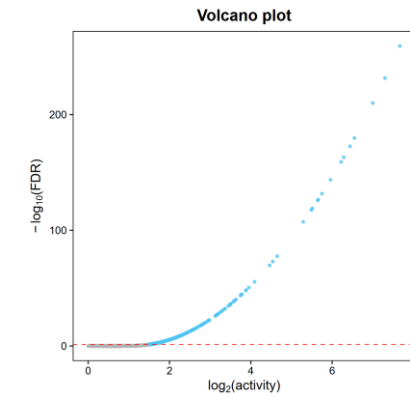
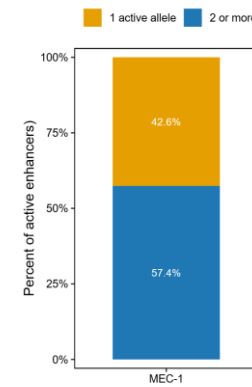
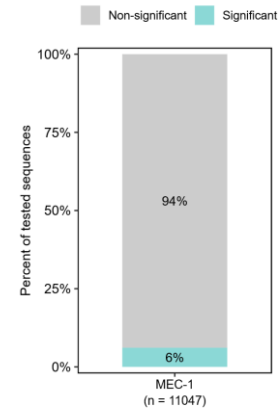


Analyse avec MPRAnalyze

2^{ème} méthode

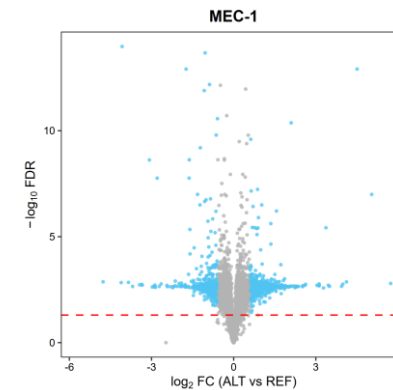
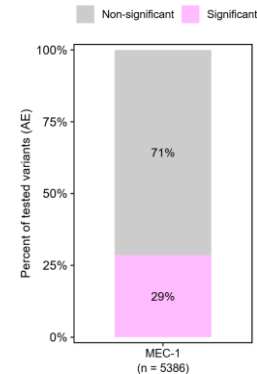
Identification des enhancers

→ 678 / 11 047 séquences avec un effet enhancer



Identification des variants régulateurs

→ 1 545 / 5 407 variants avec un effet régulateur



La majorité des variants ont un effet faible et un FDR plus élevé

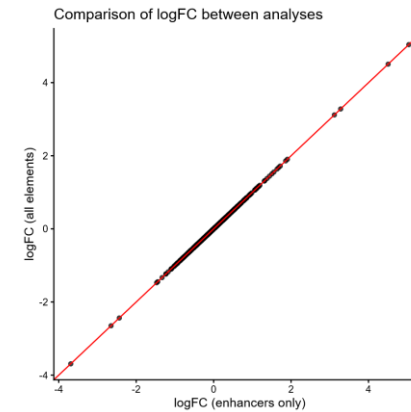
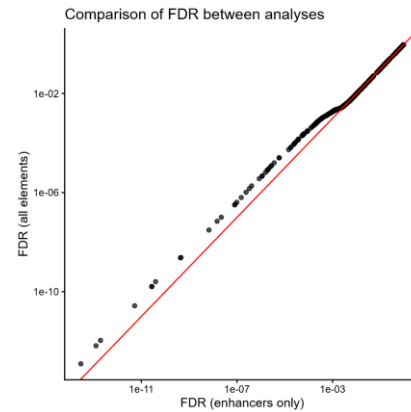
Analyse avec MPRAnalyze

1 545 variants avec un effet régulateur dont **118** sont dans des enhancers

➤ **Faut-il tester uniquement les variants dans des enhancers ?**

Tout dépend des objectifs...

Lorsqu'on compare les deux méthodes, il n'y a pas de différence entre les logFC voir très peu de différences pour les FDR



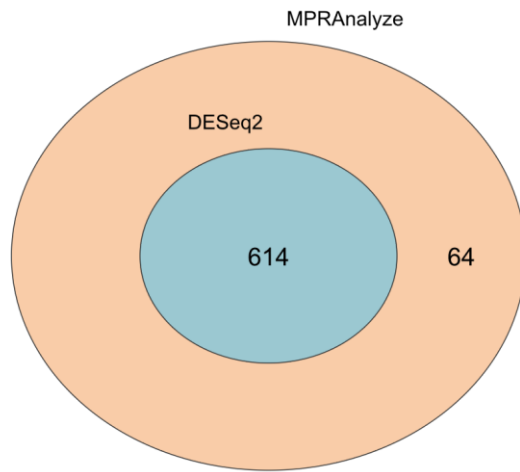
➤ **Dans quels éléments se trouvent les autres variants régulateurs ?**

- Enhancer non détecté (seuil de significativité, modèle statistique, limite taille de la séquence)
- Enhancer inactif dans le type cellulaire
- Silencer (le STARR-seq n'est pas approprié pour la détection de silencer, de même pour MPRAnalyze)

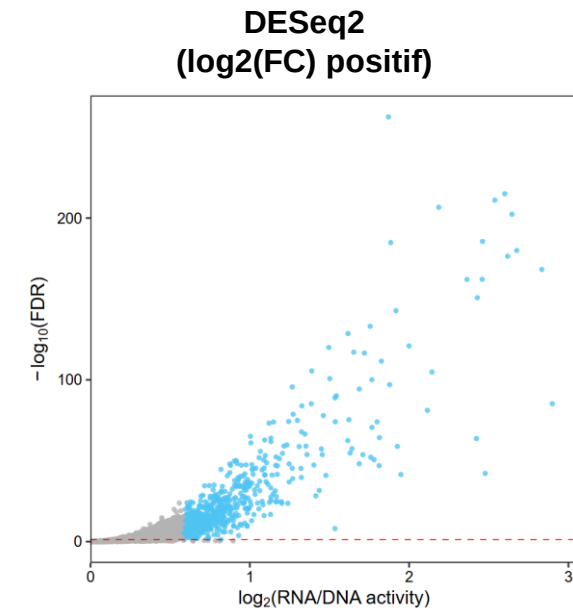
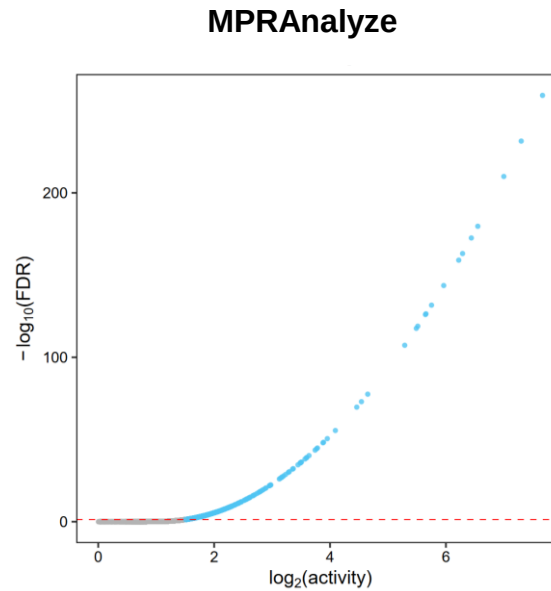
Comparaison MPRAnalyze et DESeq2

DESeq2 : analyse comparative de deux conditions ARN vs. ADN

MPRAnalyze : analyse quantitative qui prend en compte la relation ADN $\rightarrow$ ARN



MPRAnalyze identifie plus d'enhancers
que DESeq2



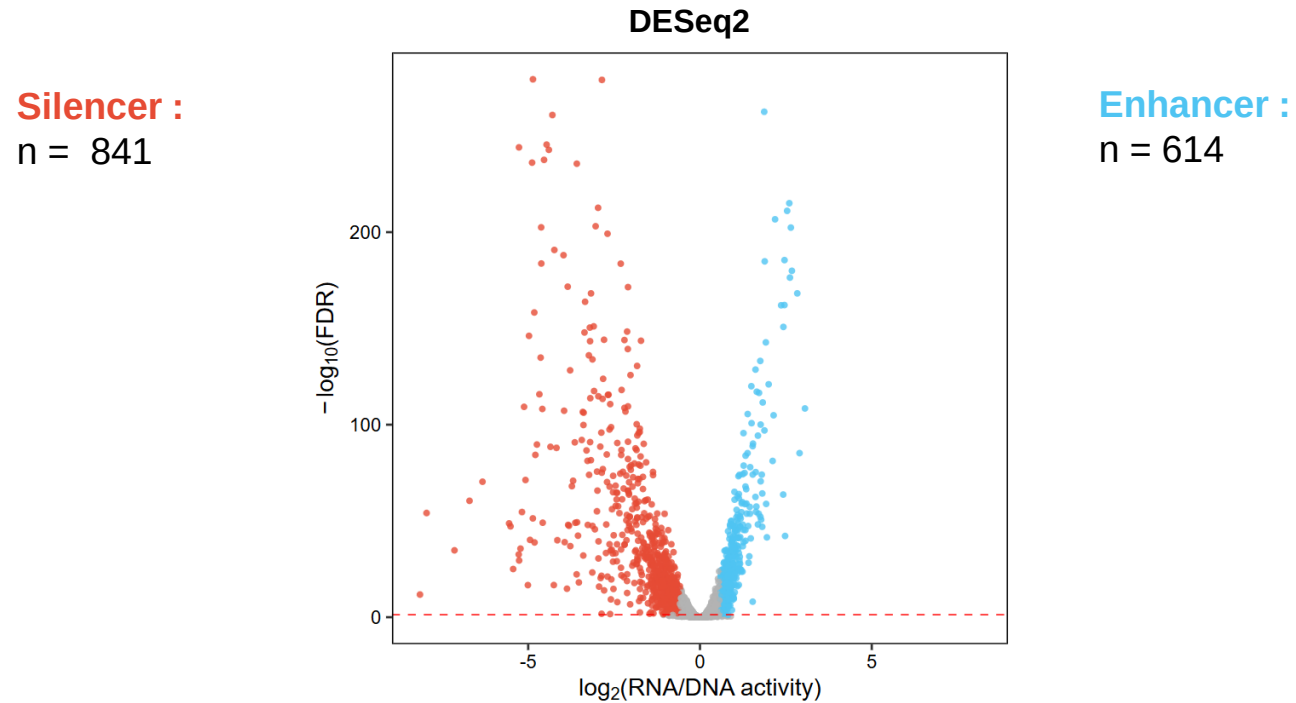
Corrélation activité : $R = 0,93$

Corrélation FDR : $R = 0,74$

Comparaison MPRAnalyze et DESeq2

DESeq2 : analyse comparative de deux conditions ARN vs. ADN

MPRAnalyze : analyse quantitative qui prend en compte la relation ADN $\rightarrow$ ARN



Méthode	Date	Expérience	Enhancer analysis	Variant analysis	Langage	Input
MPRAnalyze	2019	MPRA/STARR-seq	oui	oui (condition)	R	count
mpralm	2019	MPRA	non	oui	R	count
QuASAR-MPRA	2018	MPRA	non	oui	R	count
esMPRA	2025	MPRA	QC pour MPRAnalyze	non	Python	Fastq
CRADLE	2021	STARR-seq génomome entier	oui	non	Python	bigwig
STARRPeaker	2020	STARR-seq génomome entier	oui	non	Python	bam/bigwig

Analyses des données STARR-seq

