analyse

February 15, 2025

1 Outils

- **notebook** : un type de fichier permettant d'intercaler des section de code et des section de texte, permet aussi d'afficher les plots directements (pratique pour décrire une analyse)
- **pyvcf** : librairie python pour parser des fichier vcf et récupérer les variants qui y sont référencé
- **Biopython** : librairie de bioinformatique qui nous permet ici de parser le fichier fasta contenant toutes les ORF du virus
- **pandas** : Permet de traiter des donnés tabulaires (type csv) efficacement et facilement (accession, filtrage, trie, …) dans un objet appelé data_nf1Frame
- **seaborn** : Permet de réaliser des graphiques facilement à partir de DataFrame

2 Récupération des données

Les variants structuraux de tous les échantillons pour tous les passage ont été récupérés au préalable sur le cluster et regroupés dans un fichier csv. Chaque variants est représenté par une donné avec les entrées suivantes: - **id** : identifiant de la variation dans le fichier vcf - **svtype** : type de variant (INS, DEL, INV, DUP) - **pos** : position de début de la variation - **end** : position de fin de la variation (pour une insertion pos = end) - **svlen** : taille de la variation (négatif pour les délétions) - **alt** : séquence du variant en cas d'insertion - **dr, dv** : profondeur de read mappant sur la référence (dr) et mappant sur le variant (dv) - **depth** : profondeur de read total sur cette région du génome (dr + dv) - **af** : fréquence allélique (dv / depth : proportion de read supportant cette inversion par rapport au reads mappés sur cette région) - **sample** : échantillon d'origine (1 à 10) - **iteration** : passage d'origine (15 à 90) - **group** : les variants identiques sont regroupés avec un identifiant identique (voir ci dessous)

2.1 Groupement des variants identiques

On veut pouvoir repérer les différentes occurences d'un même variant dans différentes expériences. Pour cela les variants sont comparé entre eux en se basant sur les éléments suivants, qui doivent être identiques : - La position de début - La position de fin - La séquence alternative pour les insertions

Il a été choisis de grouper les variants seulement lorsqu'ils sont exactement identique pour être certain qu'ils ont le même impacte sur le fonctionnement biologique. Par ailleurs, cette méthode identifie de nombreuses occurences d'un même variant, ce qui montre que ce seuil est pertinent : pour 2 457 variants, 779 groupes ont été identifiés

index	pos	id	svtype	svlen	end	af	dv	dr	depth
0	1	Sniffles2.DUP.705S0	DUP	272677	272678	0.0	0	0	0
1	1	Sniffles2.DUP.3B3S0	DUP	272677	272678	0.0	0	0	0
2	1	Sniffles2.DUP.29FS0	DUP	272677	272678	0.0	0	0	0
3	1	Sniffles2.DUP.6ADS0	DUP	272677	272678	0.0	0	0	0
4	1	Sniffles2.DUP.1AS0	DUP	272677	272678	0.0	0	0	0

index	alt	sample	iteration	group	choc
0	NaN	1	15	0	cold
1	NaN	1	30	0	cold
2	NaN	1	50	0	cold
3	NaN	2	15	0	cold
4	NaN	2	30	0	cold

3 Filtrage des données

3.1 Fréquence allélique et profondeur

Pour déterminer des filtres de fréquences alléliques et de profondeur minimum qui soit pertinent on regarde quelles sont les valeurs observés dans les différents passage. Il est nécessaire de regarder tous les passages indépendament puisque la qualité globale du séquencage diffère d'un passage à l'autre. On choisiras cependant un seuil fixe pour l'ensemble des passage afin de rester consistant.

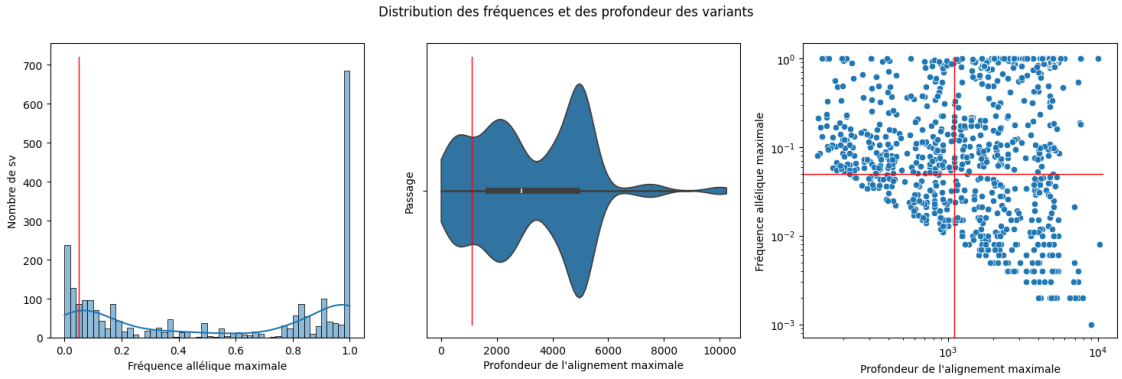
Pour les fréquences allélique, on voit sur l'histogramme on remarque le "coude" de la distribution à 0.05 (ligne rouge) : comme une grande quantité d'observations se situe en dessous de ce seuil, il est pertinent de considérer que ces observations sont les moins significatives.

Ce raisonnement est plus compliqué à appliquer sur les profondeur qui ne sont pas aussi distinctement répartis, on peut afficher des violin plots pour voir la distribution de chaque passage. La largeur du graphique représente la proportion de reads à une valeur donnée de profondeur, un boxplot est représenté à l'interieur en noir. On voit qu'avec un filtre à 1100 on écarte moins de 25% des variants, cette valeur semble pertinente.

Le dernier graphique représente chaque variant par un point positioné en fonction de sa profondeur (absysses) sa fréquence allélique (ordonnées).

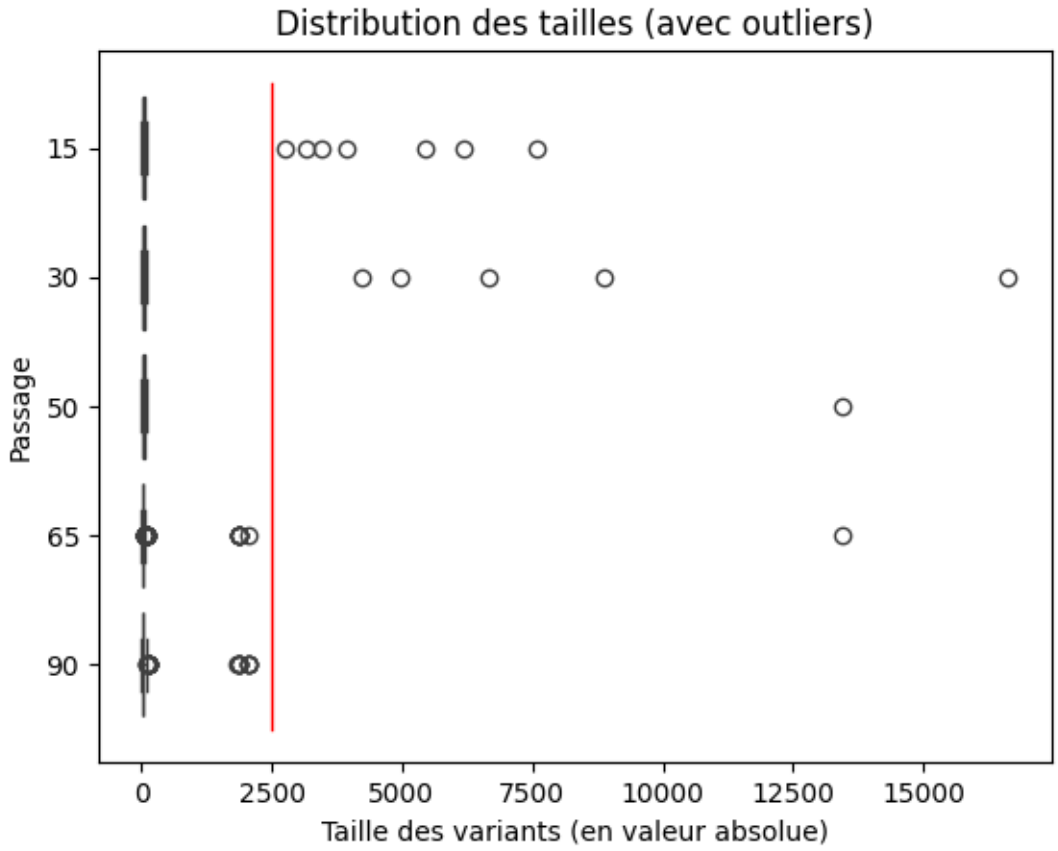
Enfin avec des seuils à 0.05 et 1100 on conservera des variants supportés par 55 reads minimum ce qui est significatif (0.05 * 1100).

Remarque : ce sont les groupes de variants (cf description plus haut) identiques qui sont filtrés et non les variants individuellement, par exemple si un variant est présent à P15 et ne remplis pas les conditions pour passer le filtre, mais qu'il est aussi présent à P50, cette fois en fréquence et en profondeur suffisante, les deux occurences seront conservé. En effet cela nous donne un information plus pertinentes sur les variants et nous permets de considérer les différents passages ensembles. On affiche donc les fréquences maximale pour chaque groupe de variants identiques.



3.2 Repérer les outliers

Des outliers sont des données inattendu qui peuvent possiblement être causé par des bugs. Dans notre cas on sait que les régions au début et à la fin du génome sont répétés, ce qui peut conduire à détecter de larges insertions dans ces régions. Afficher la distribution des longueurs de variants dans un box plot nous permet de les repérer pour les filtrer correctement.



En observant plus en détails ces donnés on remarque que soit ces variants sont positionnés au bornes du génomes, soit ce sont des régions répétés. Ci dessous tous les variants d'une taille superieur à 2 500 entre les positions 20 000 et 280 000.

On décidera donc de filtrer les variants qui ont une taille supérieur à 2 500

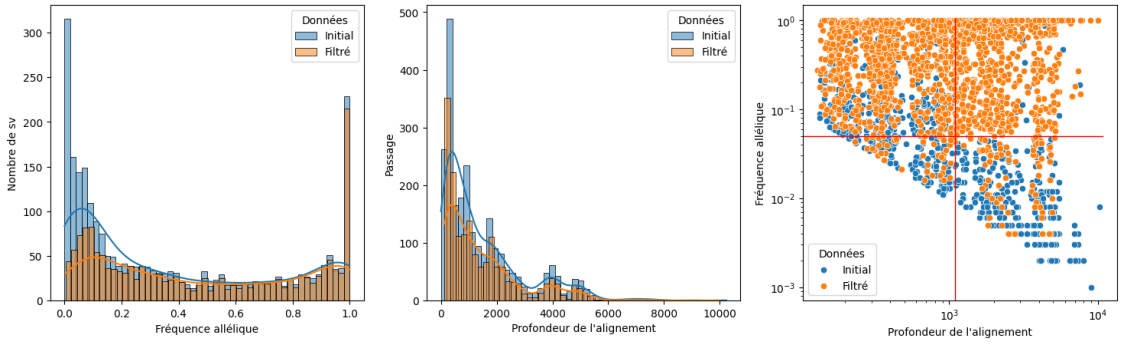
index	svlen	pos	alt
1774	16622	196282	AAAAAAAAAAAAAAAAAAAA
2018	13429	229512	NaN
2019	13429	229512	NaN
1333	8855	138818	CTCAACTGTCGTGCCGTTGA
1082	7568	85419	GGGGGGGGGGCGCTATGTGG
1447	6668	155479	CTCTGGGGCTGAGCGGGAGA
1334	6168	142608	GGGGGGGGGGGGGGGGGGGG
303	5867	22914	GAGAGAGAGAGAGAGAGAGA
1267	5459	122251	GGGGGGGGGGGGGGGGGGGG
848	4977	54238	GGGGGGGGGGGGGGGGGGGG
2203	4226	254597	AGGGAGAGCGCGCGCGCGCG
1947	3921	225969	GTGTGTGTGTGTGCGGCGGT
1930	3453	217321	GGGGGGGGGGGGGGGGGGGG
967	3157	81347	CGGTCAAGAGCTGAGACTGC
2281	2758	264795	AGCGAGGAGGAGGAGGAGAG

3.3 Impacte du filtrage

On superpose les données aux différentes étapes du filtrage pour comprendre l'impacte et voir la quantité de données conservés.

Ici j'ai choisit de montrer toutes les fréquences indépendament. On voit que même avec un filtre élevé concernant la profondeur on conserve une grande proportion des variants.

```
Proportion de variants conservés : 0.6919006919006919
```



3.4 Extraire les variants qui interfèrent avec un ORF connu

Pour cela on extrait tous les o connues du virues à partir d'un fichier FASTA. Si une variation structurelle chevauche un ou plusieurs o, on noteras lesquel.

index	id	orfs
54	Sniffles2.DEL.1D5S0	[CyHV3_ORF5_1]
668	Sniffles2.DEL.1E6S0	[CyHV3_ORF25, CyHV3_ORF26, CyHV3_ORF27]
669	Sniffles2.DEL.34BS0	[CyHV3_ORF25, CyHV3_ORF26, CyHV3_ORF27]
670	Sniffles2.DEL.2C4S0	[CyHV3_ORF25, CyHV3_ORF26, CyHV3_ORF27]
671	Sniffles2.DEL.263S0	[CyHV3_ORF25, CyHV3_ORF26, CyHV3_ORF27]

4 Exploration de P90

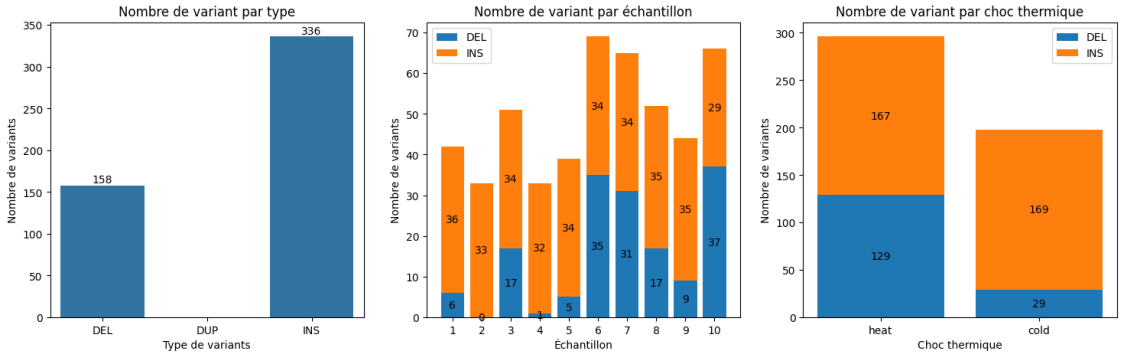
Remarque : Le code suivant peut être appliqué aux autres passages simplement en changant la variale ioi

4.1 Distributions des variants

On veut d'abord mener une analyse descriptive pour voir les différents variants qui compose les échantillons du passage P90, leur quantités, leurs positions et leur tailles.

On remarque que les échalntillons du choc chaud sont beaucoup plus riche en délétion (3ème plot) alors que le nombre d'insertions reste très stable entre les différents échantillons (2ème plot). Le nombre d'insertion étonnament proche entre le groupe chaud et froid nous laisse penser que ce pourrait être les mêmes insertions. À ce stade on peut supposer que les délétions sont plus corrélés au choc thermique que les insertions, elles ont pu être sélectionnés dasn le chaud ou éliminés dans le froid.

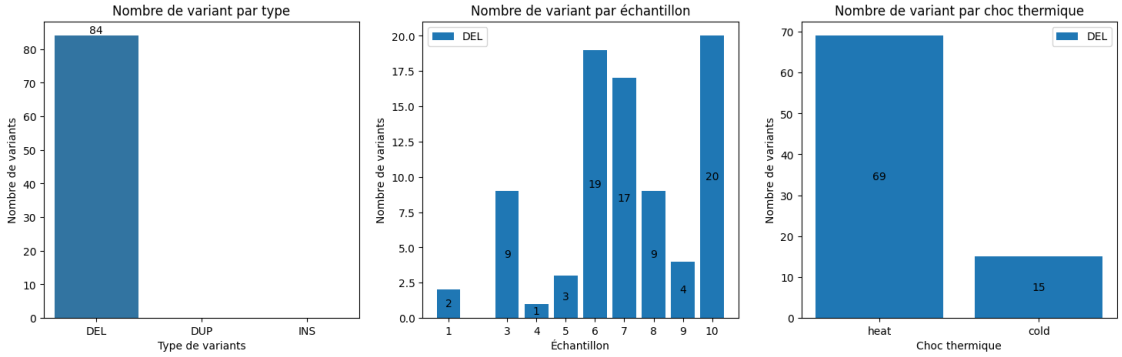
Si on suppose que la probabilité d'apparition d'une insertion et d'une délétions sont identiques (est ce que c'est vrai ?), on pourrait alors imaginer que les délétions sont contre-sélectionnés par le choc froid, donc ont un effet délétaire pour ce groupe.



Distribution des types de variants dans le passage P90

On peut répliquer l'expérience en s'intéressant uniquement aux variants présent dans des ORFs.

Les insertions qui sont présente n'ont a priori pas d'impacte sur les ORF, ce qui nous comforte dans l'hypothèse que c'est pour les délétions que le choc thermique à été discriminant.

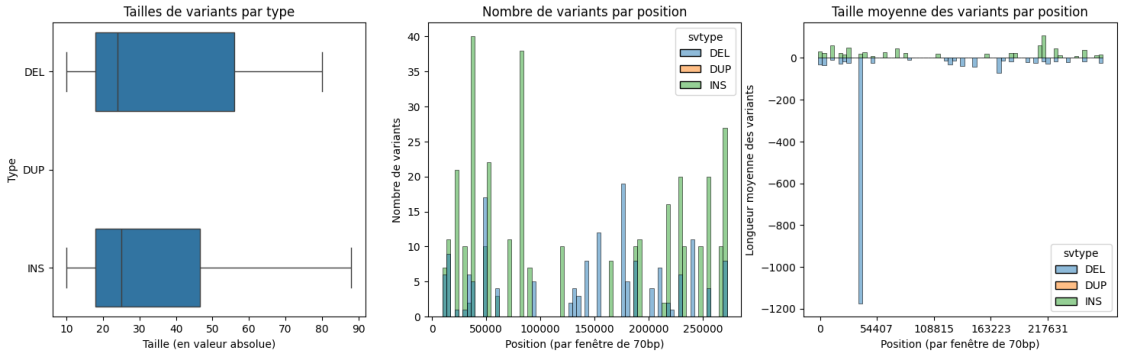


Distribution des types de variants dans le passage P90 (variants dans les ORFs uniquement)

4.2 Tailles et positions des variants

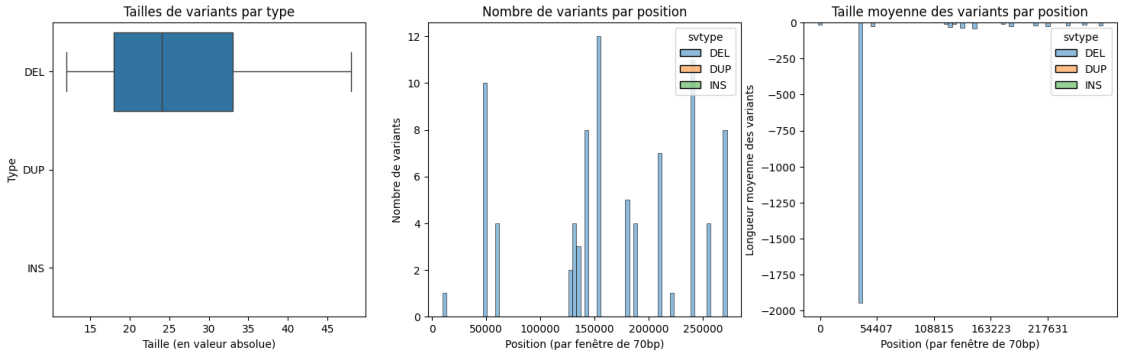
On veut analyser la taille des différents variants ainsi que leurs position sur le génome pour voir si il y a des régions qui se démarquent.

On note qu'il n'y a pas de différence notable de longueur entre les insertions et les délétions, ormis la délétion "géante" de l'orf25



Distribution des tailles et positions des variants dans le passage P90

Similairement, on peut répliquer ces graphiques sur les variants présent uniquement dans les ORFs



Distribution des tailles et positions des variants dans le passage P90 (dans les ORFs uniquement)

4.3 Comparaisons entre les échantillons

4.3.1 Méthode

On évalue la similarité entre échlantillons, paire à paire, pour voir si ceux ci partagent beaucoup de variations. On sait déjà des analyse précédentes que le groupe chaud possède nettement plus de délétions que le groupe froid. Cette analyse nous permet de voir si ces délétions sont uniques, ou si elles sont partagés entre les échantillons du groupe chaud : si on les retrouves en quantité significatives, nous pourrons affirmer avec un certain niveau de confiance que la présence/absence de ces mutations est belle est bien corrélé au choc thermique.

Soit S la matrice de similarité, la similarité entre deux échantillons i, j est donné par $S_{i,j} = Sj, i = \frac{\text{Nombre de variants présent dans } \mathbf{i \text{ et } j}}{\text{Nombre de variants présents dans } \mathbf{i \text{ ou } j}}$

Bien que cette mesure donne une bonne idée de l'homogénéité des échantillons, il pourrait être intéressant de considérer un vrai test statistique qui serait pet être plus robuste. La principale faiblesse ici est la sensibilité à la taille des échantillons : en effet si les échantillons sont trop petits, ils sont nécéssairement moins suceptible de partager des variations, d'autant que les échantillons du groupe froid présentent moins de variations que les échantillons du groupe chaud (voir plus haut).

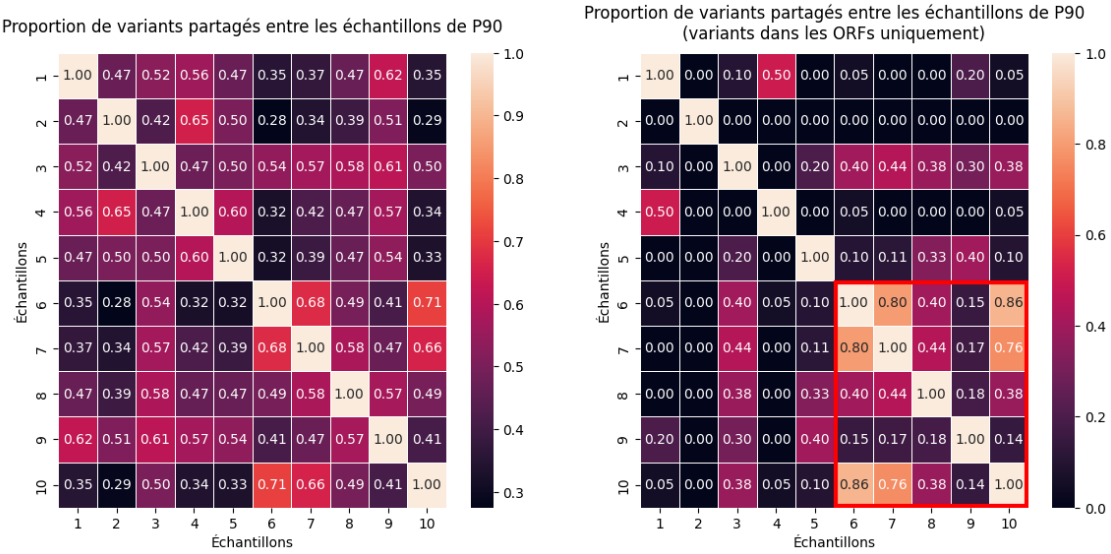
4.3.2 Résultat

Dans la 1ère matrice on observe des valeurs plus significative au sein des groupes froid / chaud, la tendance est assez similaire. Bien que les insertions présentent dans le groupe froid ne se trouvent pas dans des ORFs, l'homogénéité entre les échantillons 1 à 5 peut nous laisser penser que: - ces insertions ont un impacte sur la fitness du virus (région régulatrice ? apparition d'un ORF ? quoi d'autre ?) -> comment ont elles évolué depuis les précédents passages - ces insertions se trouvent dans des régions fortement mutagènes -> investiguer les régions de ces insertions - ces insertions sont des artéfact causés par des régions répétés -> investiguer les régions de ces insertions

Dans la seconde matrice on se concentre sur les variations présentent uniquement dans les ORFs. On remarque un cluster entre les échantillons 6 à 10, comfirmant que les délétions trouvés dans ces échantillons sont récurrentes et renforcant l'hypothèse qu'elles ont un lien avec le choc thermique. Comme expliqué précédemment, il faut faire attention à l'interprétation des similarités entre les échantillons 1 à 5 (froid) puisque ceux ci présentent très peu de mutations dans des ORFs, les

valeurs pourraientt alors être biaisé et traduire une absence de mutations dans les ORFs plutot qu'une hétérogénéité du groupe.

Enfin on remarque les échantillons 3 et 9 qui se démarquent nettement de leur groupe. Hypothèses: - un autre mécanisme entre en jeu (par exemple un autre variants, plus rare à permis de contrer le potentiel effet délétaire des délétions) - le séquencage à pu révéler des variants dans l'échantillon 3 mais pas dans les 3 autres, ce qui voudrais dire que ces variants ne serait pas nécéssairement corrélés au choc thermique, plutot un problème expérimental (échantillonage ? séquencage ?) -> regarder en détail l'échantillon 3 pour voir si il se démarque des autres (profondeur de l'alignement, distribution des variants, proportions de tags assignés, ect…) - certains adn provenant de P90-3 ont étés échangés (ou leurs tags) avec des adn provenant de P90-9 -> qui a préparé P90-3 lol ?



4.4 Recherche des variants discriminants

Si des variants sont corrélés avec le choc thermique, ils ont probablement un effet significatif dans un des deux groupes. La solution la plus direct est de rechercher les variants qui sont dans des ORFs et qui ne sont présent que dans le groupe chaud ou le groupe froid.

Cette méthode présente plusieurs inconvénients. Elle est trop restrictive et peu rater des variants présent en fréquence significativement différentes. Par exemple si un variants et présent 1 fois dans le chaud et 5 fois dans le froid il ne sera pas détecté malgrès une différence significative, alors que si il est présent 0 fois dans le chaud et 1 fois dans le froid, il sera détecté (situation beaucoup moins significative).

Il faudrais alors mettre en place un test statistique plus robuste qui nous permette de dire si l'apparition d'une variation est corrélé avec le choc thermique (fisher ? khi2 ? …).

Suite à cela nous avons choisis d'investiguer le variant de l'ORF78 qui semble être la plus importante.

```
                          id     pos      af  \
index
```

```
54     Sniffles2.DEL.1D5S0    9457  0.594
733    Sniffles2.DEL.2D2S0   47346  0.490
877    Sniffles2.DEL.29CS0   60281  1.000
880    Sniffles2.DEL.2E3S0   60281  1.000
883    Sniffles2.DEL.329S0   60281  1.000
887    Sniffles2.DEL.274S0   60281  1.000
1277   Sniffles2.DEL.2F2S0  127195  0.070
1284   Sniffles2.DEL.32ES0  127195  0.054
1318   Sniffles2.DEL.302S0  133271  0.857
1320   Sniffles2.DEL.332S0  133271  0.783
1325   Sniffles2.DEL.33DS0  133271  0.840
1345   Sniffles2.DEL.318S0  143509  0.905
1347   Sniffles2.DEL.349S0  143509  0.852
1352   Sniffles2.DEL.35FS0  143509  0.891
1403   Sniffles2.DEL.323S0  152722  0.073
1404   Sniffles2.DEL.357S0  152722  0.067
1406   Sniffles2.DEL.379S0  152722  0.050

                                                 orfs
index
54                                    [CyHV3_ORF5_1]
733    [CyHV3_ORF25, CyHV3_ORF26, CyHV3_ORF27]
877                                    [CyHV3_ORF40]
880                                    [CyHV3_ORF40]
883                                    [CyHV3_ORF40]
887                                    [CyHV3_ORF40]
1277                                   [CyHV3_ORF68]
1284                                   [CyHV3_ORF68]
1318                                   [CyHV3_ORF69]
1320                                   [CyHV3_ORF69]
1325                                   [CyHV3_ORF69]
1345                                   [CyHV3_ORF78]
1347                                   [CyHV3_ORF78]
1352                                   [CyHV3_ORF78]
1403                                   [CyHV3_ORF82]
1404                                   [CyHV3_ORF82]
1406                                   [CyHV3_ORF82]
```

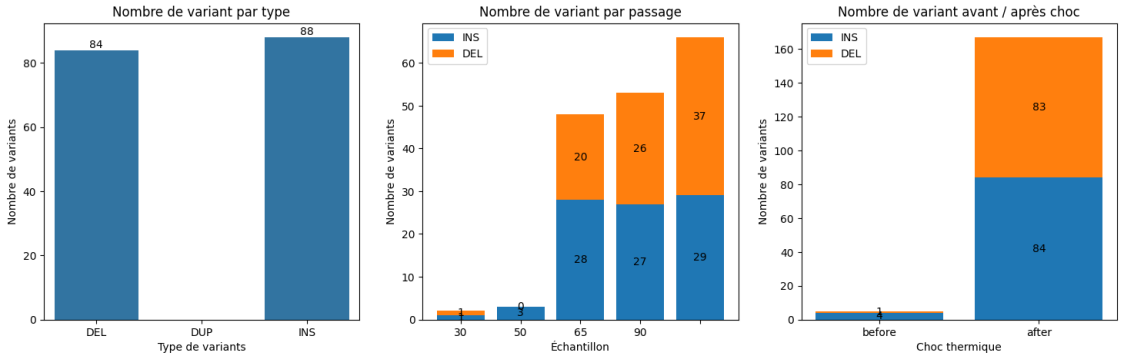
5 Analyse de l'échantillon 10

Remarque : Cette analyse peut être lancé sur n'importe quel échantillon en modifiant la variable **soi** ci dessous.

On veux comparer les différent passages d'un même échantillon pour voir si il y a un évolution particulière dans la distribution des variants. C'est quasi identique à l'anlayse précédente, on compare les passage d'un échantillon donné au lieu de comparer les échantillons d'un passage donc je ne vais pas trop rentrer dans les détails

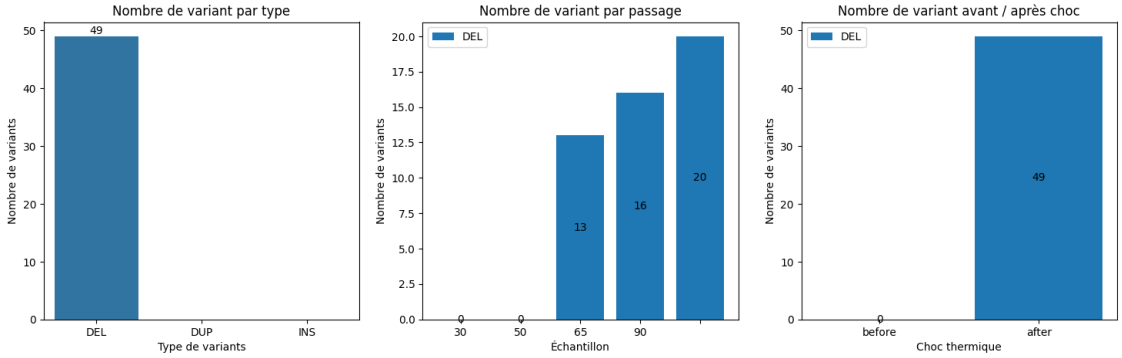
5.1 Distribution des variants

On compte le nombre de variant par type dans tous les passages de l'échantillon 10 (P15-10 à P90-10). Sur le 2ème plot on compte par passage, et sur le 3ème plot on distingue avant (< P30) et après (>= P30) le choc.



Distribution des types de variants dans les passages de l'échantillon 10

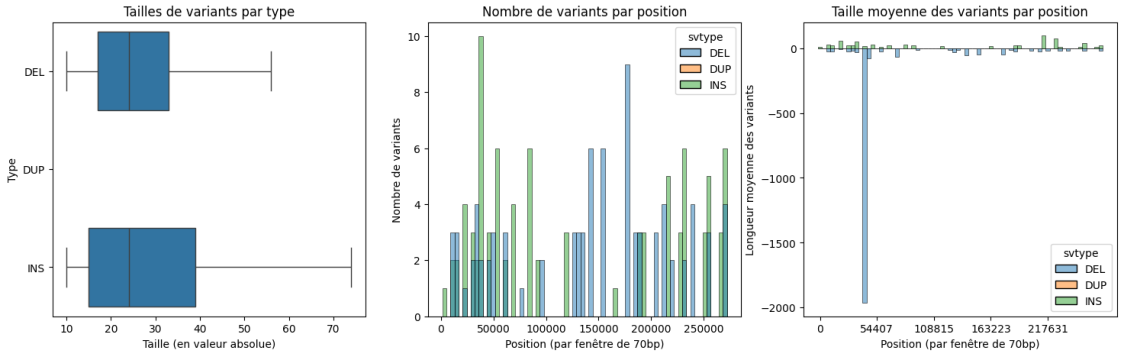
On réitère avec les variant qui chevauchent des ORFs



Distribution des types de variants dans les passages de l'échantillon 10 (variants dans les ORFs uniquement)

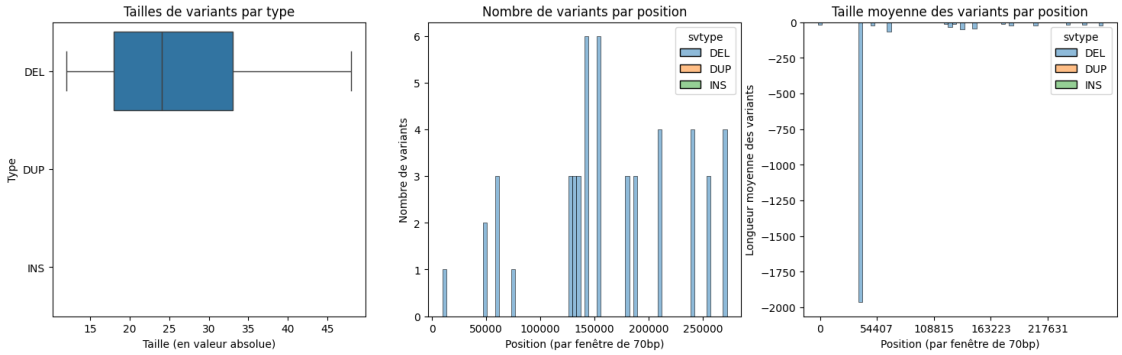
5.2 Tailles et positions des variants

Dans l'ensemble des passages de l'échantillon 10, on mesure la taille des variants et leur répartition le long du génome.



Distribution des tailles et positions des variants dans l'échantillon 10

On réitère en filtrant les variants dans des ORFs

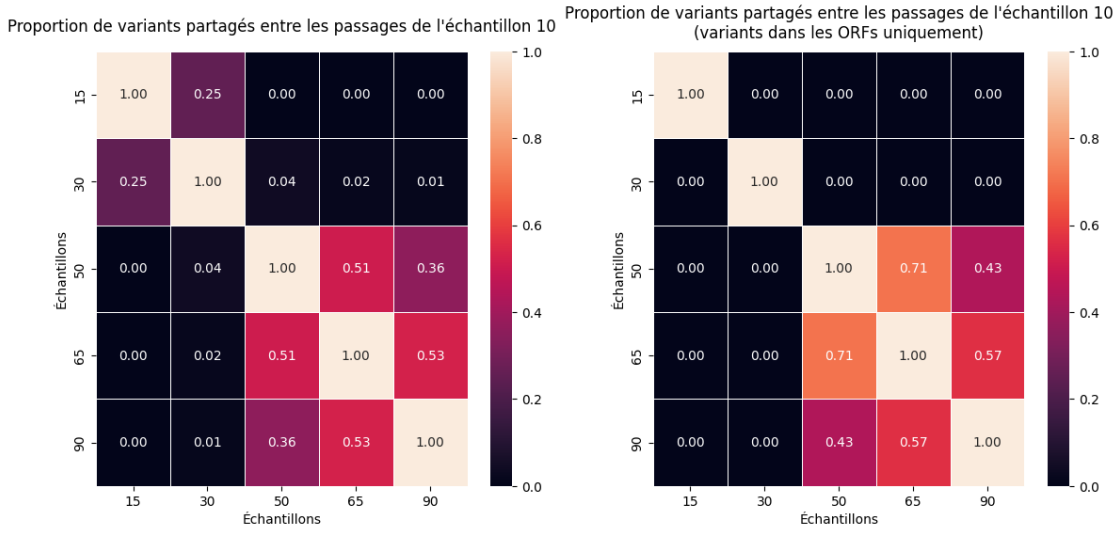


Distribution des tailles et positions des variants dans l'échantillon 10 (variants dans les ORFs uniquement)

5.3 Similarité entre les passages

Mesure de similarité par paire de passage.

On remarque bien ici une différence entre avant le choc et après le choc (cependant on voit précedemment qu'il y a très peu de variants avants le choc donc attention aux potentiel biais)



6 Perspectives

- test d'homogénéité formel pour comparer deux ensembles de variants
- test de corrélation formel pour déterminer si:
  - l'apparition d'un variant est corrélé avec le choc thermique -> on peut soit comparer les échantillons 1-5 vs 6-10 dans chaque passage, soit comparer le passage 30 vs les autres pour chaque échantillons, ...
  - la corrélation entre la présence absence de 2 variants pour déterminer si ils ont des effets complémentaires
- il s'agit de pouvoir réaliser l'analyse qui à ici été faite graphiquement, de facon statistique avec des métrique plus facilement interprétables
- investiguer échantillons 3 et 9
- considérer un seuil de similarité entre échantillon plus faible pour les considérer identiques (100% pour le moment)
- le choc thermique peut il être à la source de l'apparition / la disparition d'un variant ? plus dans le chaud ou dans le froid ?