Application des outils à des données réelles

Indice de Diversité de Hill

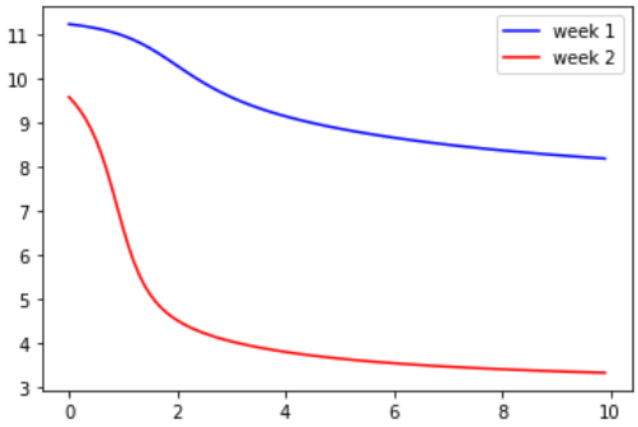
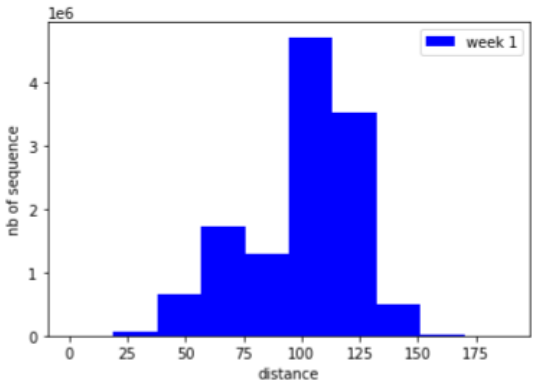


FIGURE 1 – Evolution de l'indice de diversité de Hill en fonction de la valeur de α

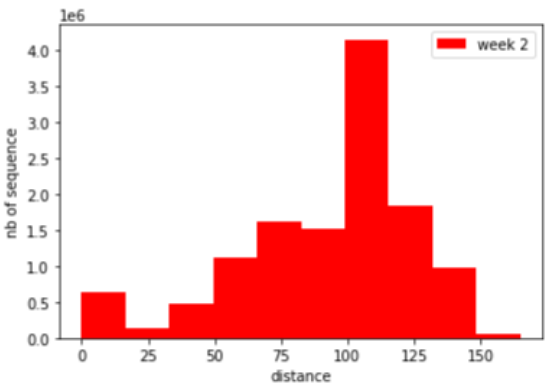
Nous observons que le répertoire de la deuxième semaine possède des valeurs de diversité de Hill inférieure à celles du répertoire de la première semaine. On peut donc supposer qu'il y a plus de diversité dans le répertoire de la semaine 2.

Pairwise Distance Distribution

Le calcul du vecteur de Pairwise Distance Distribution requiert beaucoup de temps d'éxécution, nous avons donc fait le choix d'échantilloner les répertoires et calculer la Pairwise Distance Distribution sur ces échantillons. Nous avons choisi des échantillons de 10000 séquences pour obtenir un temps de calcul raisonnable et en se basant dur les paramètres du package R SumRep.
Nous avons effectué 3 tests en échantillonant aléatoirement 10000 séquences dans les répertoires. On obtient ainsi des distributions similaires selon l'échantillon considéré.
Sur la courbe du répertoire de la semaine 2, on remarque un pic à 0 et un pic à environ 10 ce qui montre que beaucoup de séquence sont similaires dans ce répertoire.

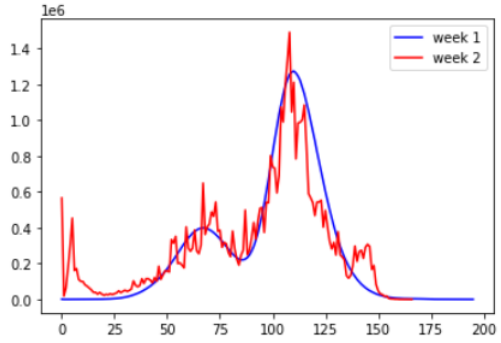


Semaine 1

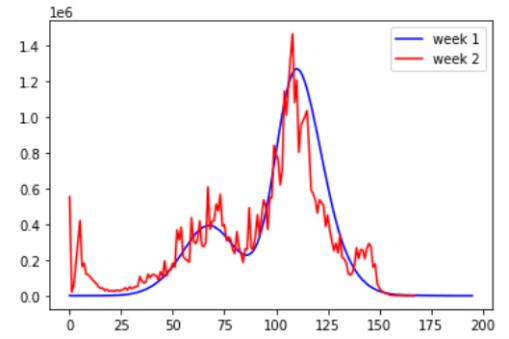


Semaine 2

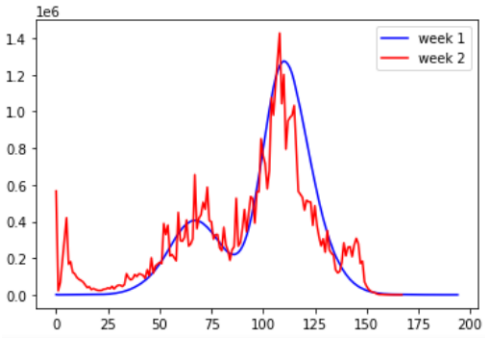
FIGURE 2 – Pairwise Distance Distribution sous forme d'histogramme sur un échantillon de 5000 séquences.



Echantillon 1



Echantillon 2



Echantillon 3

FIGURE 3 – Pairwise Distance Distribution pour 3 échantillons diffrents

Mise en Réseau

La mise en réseau d'un répertoire d'une très grande taille nécéssite un temps de calcul conséquent. Nous avons donc fait le choix déchantilloner aléatoirement 5000 séquences pour chaque répertoire, afin de réduire le temps d'éxécution.
Chaque nœud représente une séquence et la taille de ce nœud est proportionnelle au nombre de séquences identiques à celle-ci. Un arc relie deux nœuds si la distance de Levenshtein entre les deux séquences est inférieure à un seuil.
Nous avons effectué deux tests pour deux valeurs de seuils différentes : 20 et 40.
Concernant les graphes avec une valeur seuil de 20, on remarque que pour le répertoire de la semaine 1 les nœuds sont de petite taille et sont dispersés. On observe tout de même des petits groupes de séquences. Dans le graphe du répertoire de la semaine 2 on observe des nœuds de grande taille.
Concernant les graphes avec une valeur seuil de 40, pour le répertoire de la semaine 1 on remarque que les nœuds sont toujours de petite taille cependant la taille des clusters est plus grande. Pour le graphe du répertoire de la semaine 2, on observe des nœuds de grande taille et également des clusters de plus grande taille.

Seuil à 20

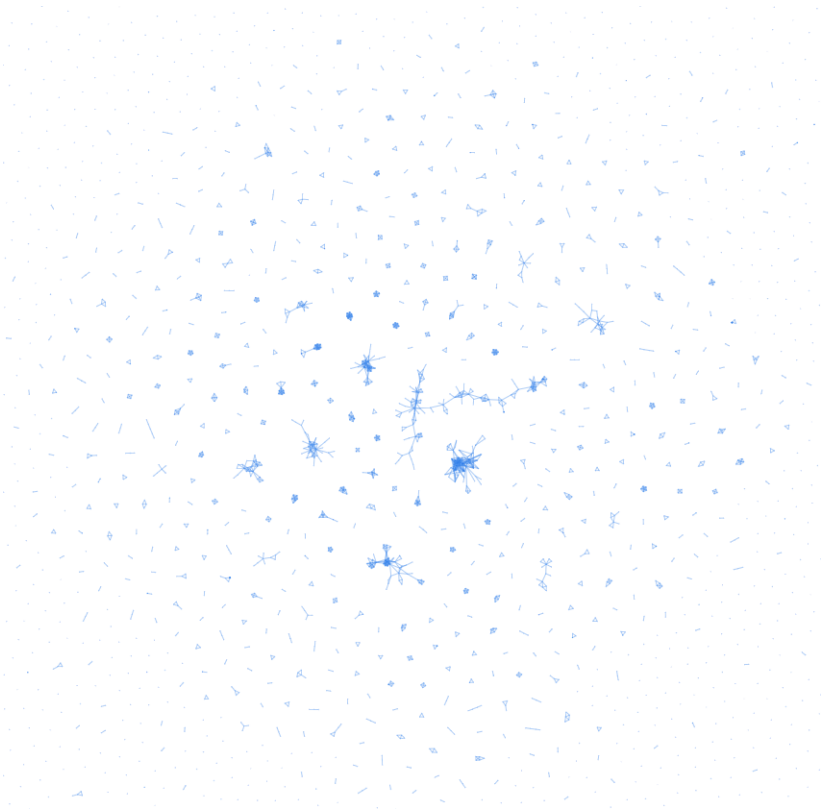


FIGURE 4 – Réseau des séquences du répertoire de la semaine 1 sur un échantillon de 5000 séquences et une valeur seuil fixée à 20

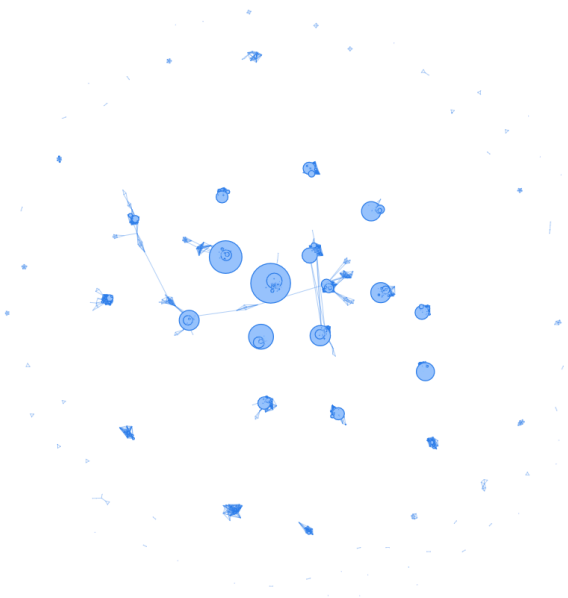


FIGURE 5 – Réseau des séquences du répertoire de la semaine 2 sur un échantillon de 5000 séquences et une valeur seuil fixée à 20

Seuil à 40

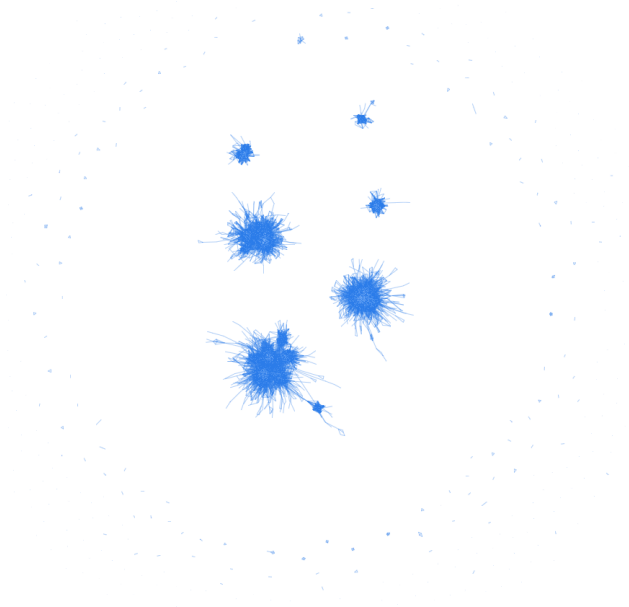


FIGURE 6 – Réseau des séquences du répertoire de la semaine 1 sur un échantillon de 5000 séquences et une valeur seuil fixée à 40

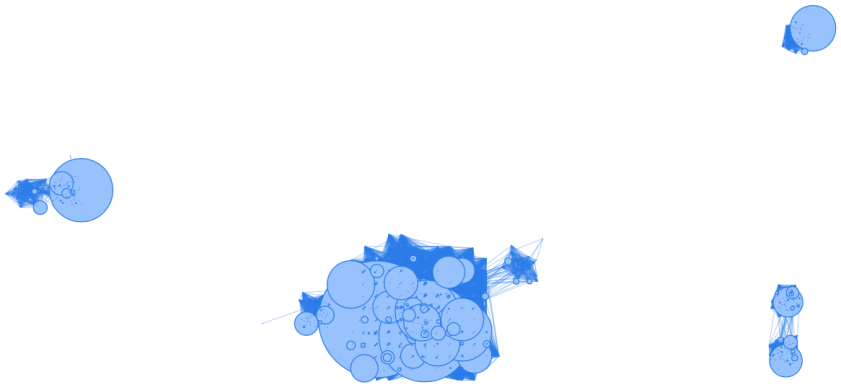


FIGURE 7 – Réseau des séquences du répertoire de la semaine 2 sur un échantillon de 5000 séquences et une valeur seuil fixée à 40

Le temps d'affichage pour un graphe de cette taille étant assez long, on peut se comparer les caractéristiques de chaque graphe comme le nombre de nœds et d'arêtes, le degré des nœuds et le nombre de composantes connexes.

```
Informations du graphe du répertoire de la semaine 1 :

Graph with 991 nodes and 117 edges

Average degree :  0.11806256306760847

Number of connected components :  885

-----------------------------------------------------------
Informations du graphe du répertoire de la semaine 2 :

Graph with 454 nodes and 4023 edges

Average degree :  8.861233480176212

Number of connected components :  59

-----------------------------------------------------------
```

FIGURE 8 – Caractéristiques des deux graphes.

Coefficient de corrélation de Pearson

	week1	week2
week1	1.000000	0.866621
week2	0.866621	1.000000

FIGURE 9 – Coefficient de corrélation de Pearson entre les vecteurs de diversité de Hill des deux répertoires

Jensen-Shannon Divergence

Le calcul de la Jensen-Shannon Divergence s'effectue avec les vecteurs de Pairwise Distance Distribution. On approxime donc toujours les résultats en échantillonant 10000 séquences pour chaque répertoire.
Nous obtenons une valeur très proche de 0, ce qui montre que les distributions sont très similaires. En effet, sur la figure 2, on voit que les courbes suivent la même allure. La faible différence s'explique par l'irrégularité de la distribution du répertoire de la semaine 2.

```
JSD value :  0.03653448184683998
```

FIGURE 10 – Valeur JSD entre les deux répertoires

Pourcentage de Similarité

Nous obtenons un pourcentage de similarité de 0%. En effet, nous nous attendions à avoir un pourcentage de similarité élevé étant donné que les répertoires proviennent de la même personne. Cependant, nous avons appris que les répertoires d'un individus peuvent contenir théoriquement jusqu'à 10^{12} séquences différentes. De ce fait, un prélèvment ne donnera qu'environ 10^5 de ces séquences, ce qui représente une petite partie de l'ensemble réel. Ainsi, si on effectue un deuxième préleèvement il y a très peu de chances de tomber sur les mêmes séquences, ce qui explique le taux de similarité de 0%.

VDJ Usage

En répertoriant les gènes V, D et J de chaque séquences dans chaque répertoire ainsi que leur proportion, on peut estimer la diversité de ceux-ci.

On peut trace un histogramme pour visualiser les gènes V et J utilisés dans chaque répertoire. Pour la visualisation, nous ne prenons en compte que des gènes qui apparaissent plus de 100 fois dans le répertoire. (Voir Figure 9 et Figure 10).
On peut également comparer la répartition des gènes en commun pour les deux répertoires. Dans ce cas, on considère uniquement les gènes V et J en commun. (Voir Figure 11).

Conclusion

Avec les résultats obtenus, nous avons déterminé les caractéristiques des deux répertoires et nous pouvons donc les comparer pour constater l'évolution du répertoire.

Nous pouvons supposer que le répertoire de la semaine 2 est plus diversifié que celui de la semaine 1. Tout d'abord, les indices de diversité de Hill pour le répertoire de la semaine 2 sont trés inférieurs à ceux de répertoire de la semaine 1. De plus, avec la Pairwise Distance Distribution, on voit que beaucoup de séquences sont proches lors de la deuxième semaine. On peut donc supposer que le vaccin à entrainé la production du même type d'anticorps pour combattre le virus, ce qui conduit à l'augmentation de la diversité immunitaire chez l'individu. Cette hypothèse est reforcée par le fait que dans les graphes obtenus, on observe des nœuds de grande taille ce qui montre que plusieurs exemplaires d'une même séquence sont présents dans le répertoires.

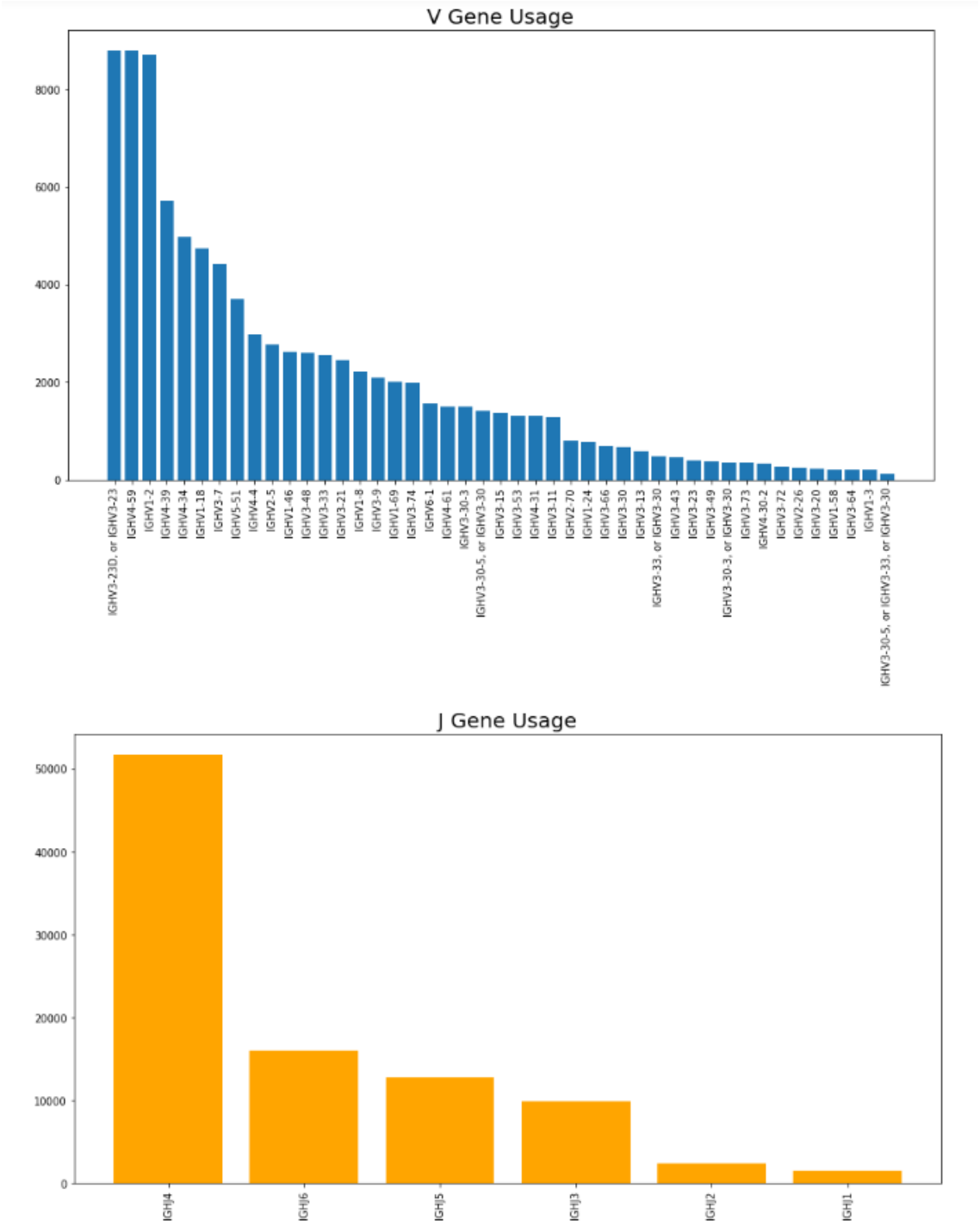


FIGURE 11 – Répartition des gènes V et J dans le répertoire de la semaine 1.

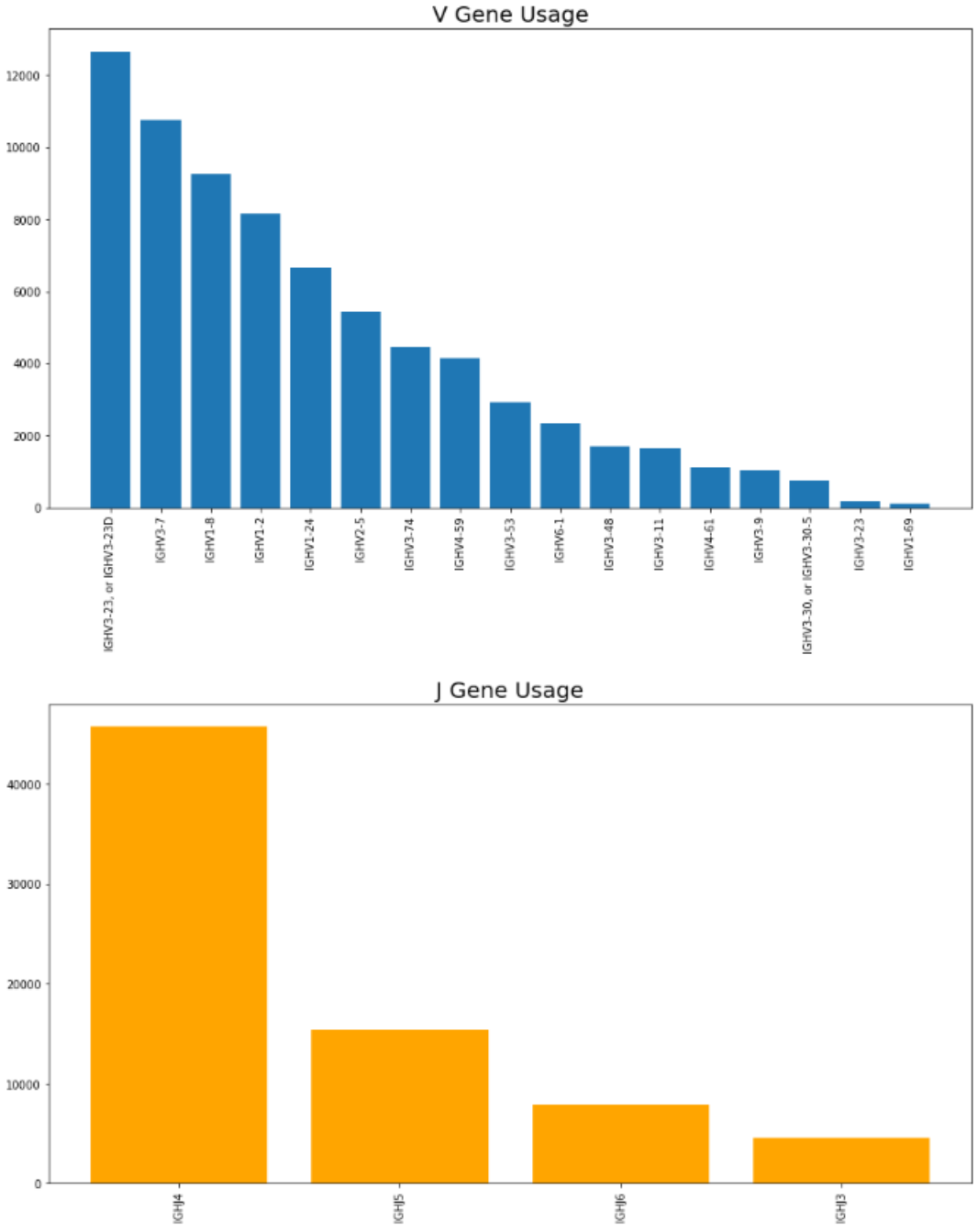


FIGURE 12 – Répartition des gènes V et J dans le répertoire de la semaine 2.

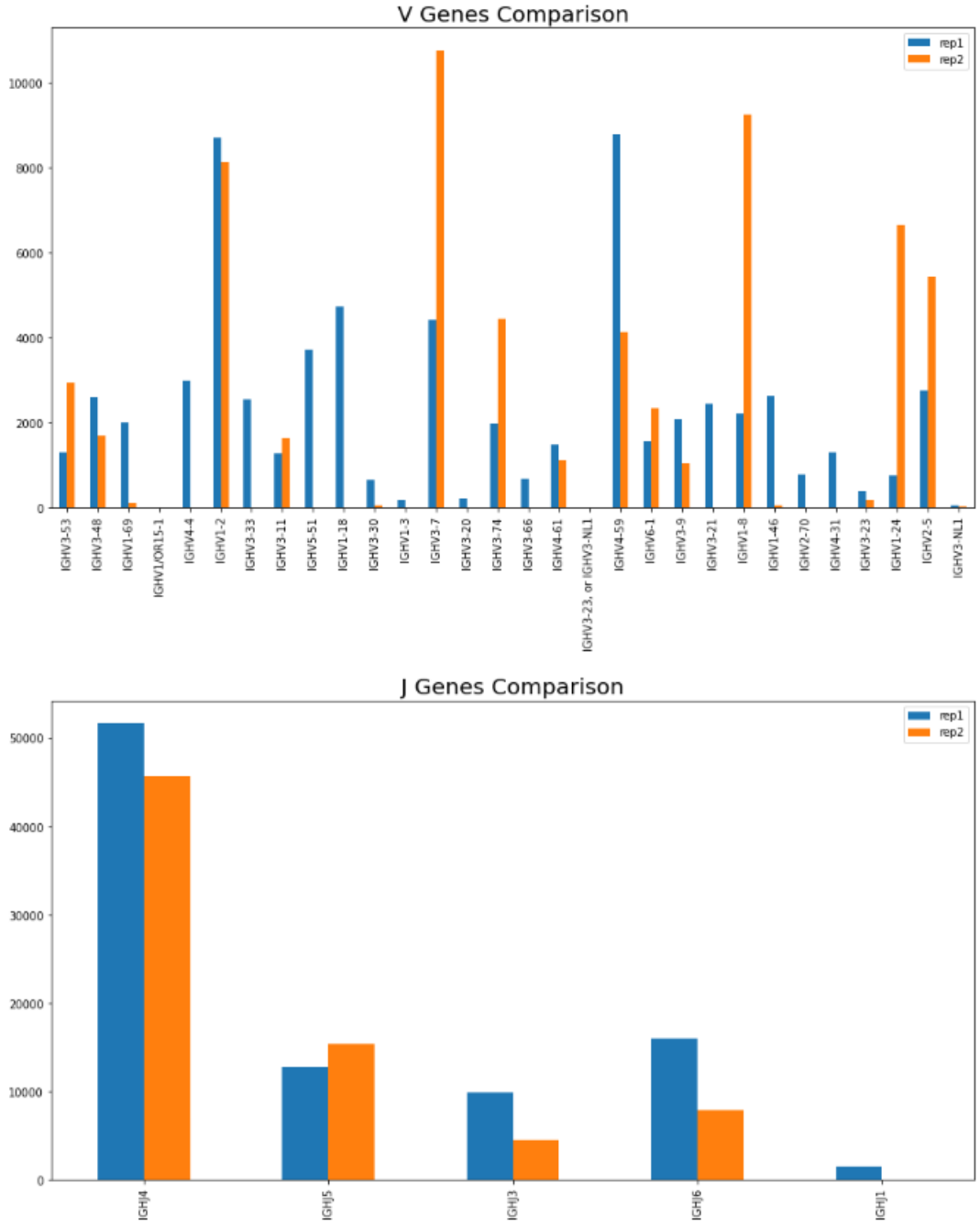


FIGURE 13 – Comparaison de la répertition des gènes V et J en commun pour les deux répertoires.