# Outils Disponibles

## RDI : Repertoire Dissimilarity Index

L'outil RDI est un processus permettant de calculer des distances entre différents répertoires. Cette méthode est utile pour déterminer si des répertoires sont éloignés ou encore détecter des groupes de répertoires qui sont similaires.

Cette méthode prend en entrée une population ainsi que l'appartenance de chacun des individus à un ou plusieurs répertoire et renvoie une matrice de distance entre chaque répertoires.

Pour obtenir ces valeurs, on procède en plusieurs étapes. Tout d'abord, RDI va ajuster la taille des répertoires à la taille du plus petit répertoire en les échantillonant aléatoirement. Ensuite, on compte les occurences des caractéristiques étudiées. Les fréquences sont normalisées et converties en probabilités. Les distances sont ensuite calculées selon la méthode RMSD (Root Mean Square Deviation). On répète cette procédure plusieurs fois en échantillonant aléatoirement à chaque itération.
La valeur RDI est obtenu en effectuant la moyenne des valeurs calculées.

## DivE : Diversity Estimator

L'outil Diversity Estimator (DivE) est une méthode heuristique pour estimer le nombre d'espèces en fonction du nombre d'individus.

En effet, un échantillon d'une population ne représente pas toute la diversité de la population. La résolution de ce problème d'espèces non-observées est importante dans le domaine de l'immunologie. L'application de cette méthode à une population de lymphocytes permet de mieux comprendre les réponses immunitaires et permet d' estimer la diversité de cette population.

L'estimateur DivE s'appuie sur plusieurs modèles mathématiques, fournis avec la package R, pour prédire la croissance du nombre d'espèces. Ces modèles sont appliqués à des échantillons de la population donnée et sont classés selon plusieurs critères :

- Divergence : Erreur moyenne entre les données et les prédictions. Le modèle doit prédire les données sur lesquels on l'a appliqué.
- Précision : A partir d'un sous-échantillon, le modèle doit prédire correctement l'évolution du reste de l'échantillon.
- Similarité : Le modèle doit effectuer des prédictions similaires à partir de n'importe quel sous-échantillon.
- Plausibilité : Le nombre d'espèce ne doit pas décroitre et le taux d'apparition d'espèces ne doit pas augmenter. Autrement dit, on doit avoir $\forall x \geq 1,\ S'(x) \geq 0 \wedge S''(x) \leq 0$, où $S$ est la fonction qui associe le nombre d'espèce au nombre d'individus.

Grâce à ces critères, on associe un score à chacun des modèles. On sélectionne les cinq meilleurs modèles et on effectue la moyenne géométrique de leur prédiction (nombre d'espèces sur la population totale) pour donner une bonne estimation.

## ReCoLD : Repertoire Comparison in Low Dimension

ReCoLD est une méthode pour comparer différents répertoires de récepteurs de lymphocytes T.

Cet outil prend en entrée un ensemble de séquence de récepteurs de lymphocytes T à étudier. La méthode consiste en plusieurs étapes :

- Calculer la matrice de dissimilarité entre les séquences à l'aide de l'algorithme Smith-Waterman et d'une matrice de score.
- Utiliser des méthodes de réduction de dimension en préservant les relations inter-séquences.
- Estimer la distribution des séquences avec la méthode d'estimation par noyaux.
- Etablir des clusters d'échantillon en calculant la valeur de diversité Jensen-Shannon (JSD) entre les probabilités calculées.
- Identifier les séquences qui contribuent le plus aux différences. Il s'agit des 1% de séquences qui ont les valeurs JSD les plus élevées. Pour ces séquences, on calcule alors les fréquences des acides aminés.

## ImmuneRef

ImmuneRef est un outil sous forme de package R permettant d'analyser un répertoire immunitaire en donnant 6 mesures de diversité interprétables :

— L'indice de diversité de Hill est calculé pour $a \in [0, 10]$.

— Les fréquences d'acides aminés par positions sont calculées dans chaque séquences

— Le nombres de $k$-mers, contenant $m$ gaps ou moins de $m$ gaps, dans chaque séquences.

— Nombre de gènes de lignée germinale dans chaque répertoire. Les gènes de lignée germinale (germline genes) sont défins dans la base de donnée International ImMunoGeneTics.

— Calcul du chevauchement clonal entre répertoires avec la formule :

$$overlap(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|}$$

où $X$ et $Y$ sont deux répertoires.
Cette valeur représente la similarité entre paires de répertoires en prenant en compte le clonage.

— Etablissement d'un réseau où chaque nœuds représente une séquence de récepteur, connectés par des arêtes de calculées avec la distance de Levenshtein.

## SumRep

SumRep est un outil sous forme de package R qui permet de calculer plusieurs valeurs statistiques concernant un répertoire immunitaire. Cet outil prend en entrée une base de donnée de séquences récepteurs de lymphocytes B et T sous format AIRR.

SumRep peut servir à différentes choses, comme détecter des motifs dans un alignement de séquences, ou encore des réarrangements entre séquences. SumRep est égaement utile pour comparer deux répertoires. Par exemple, la fonction *compareRepertoires* prend deux répertoires en argument et retourne plusieurs valeurs qui permettent de les comparer, comme par exemple leur différence longueur de séquence. SumRep dispose d'outils graphique permettant de visualiser ces valeurs au sein d'un répertoire ou entre deux répertoires.
On peut aussi utiliser SumRep pour générer des données simulées à partir de données expérimentales.
La liste complète des valeurs statistiques supportées par SumRep est visible en Figure 1.

| Summary statistic | Annotations | Clustering | Phylogeny | Implementation |
|---|---|---|---|---|
| Pairwise distance distribution | No | No | No | `stringdist` (25) |
| *k*th nearest neighbor distribution | No | No | No | `stringdist` |
| GC-content distribution | No | No | No | `ape` (26) |
| Hotspot motif count distribution | No | No | No | `Biostrings` (27) |
| Coldspot motif count distribution | No | No | No | `Biostrings` (27) |
| CDR3 length distribution | Yes | No | No | Tool-provided |
| Joint distribution of germline gene use | Yes | No | No | `sumrep` |
| Pairwise CDR3 distance distribution | Yes | No | No | `stringdist` |
| Atchley factor distributions | Yes | No | No | `HDMD` (28) |
| Kidera factor distributions | Yes | No | No | `Peptides` (28) |
| Aliphatic index distribution | Yes | No | No | `Peptides` |
| G.R.A.V.Y. index distribution | Yes | No | No | `alakazam` (21) |
| Polarity distribution | Yes | No | No | `alakazam` |
| Charge distribution | Yes | No | No | `alakazam` |
| Basicity distribution | Yes | No | No | `alakazam` |
| Acidity distribution | Yes | No | No | `alakazam` |
| Aromaticity distribution | Yes | No | No | `alakazam` |
| Bulkiness distribution | Yes | No | No | `alakazam` |
| Per-gene substitution rate | Yes | No | No | Tool-provided + `sumrep` |
| Per-gene-per-position substitution rate | Yes | No | No | Tool-provided + `sumrep` |
| Per-base substitution model | Yes | No | No | `shazam` (21) |
| Per-base mutability model | Yes | No | No | `shazam` |
| Positional distance between mutations distribution | Yes | No | No | `sumrep` |
| Distance from germline to sequence distribution | Yes | No | No | `stringdist` |
| V gene 3′ deletion length distribution | Yes | No | No | Tool-provided |
| V gene 5′ deletion length distribution | Yes | No | No | Tool-provided |
| D gene 3′ deletion length distribution | Yes | No | No | Tool-provided |
| D gene 5′ deletion length distribution | Yes | No | No | Tool-provided |
| J gene 3′ deletion length distribution | Yes | No | No | Tool-provided |
| J gene 5′ deletion length distribution | Yes | No | No | Tool-provided |
| VD (or VJ) insertion length distribution | Yes | No | No | Tool-provided |
| DJ insertion length distribution | Yes | No | No | Tool-provided |
| VD (or VJ) insertion transition matrix | Yes | No | No | `sumrep` |
| DJ insertion transition matrix | Yes | No | No | `sumrep` |
| V/J in-frame percentage | Yes | No | No | Tool-provided + `sumrep` |
| Cluster size distribution | Yes | Yes | No | Custom |
| Hill numbers (diversity indices) | Yes | Yes | No | `alakazam` |
| Selection estimates (using the BASELINe method) | Yes | Yes | No | `shazam` |
| Sackin index distribution | Yes | Yes | Yes | `CollessLike` (29) |
| Colless-like index distribution | Yes | Yes | Yes | `CollessLike` |
| Cophenetic index distribution | Yes | Yes | Yes | `CollessLike` |

*Annotation denotes whether annotation of the V(D)J germline segment is required, Clustering denotes whether clonal clustering is required, and Phylogeny denotes whether lineage tree inference is required. "Tool-provided" means that the summary can be directly computed from the output of an annotation tool; for example, the CDR3 length distribution is exactly the frequencies of values in the* `junction` *column of the annotated dataset. Per-gene substitution rate is defined to be the number of observed mutations in sequences assigned to that gene, in the segment of the sequence assigned to that gene's region, divided by the length of the segment. Per-gene-per-position substitution rate is similarly defined, but separately computed for each position in the sequence.*

FIGURE 1 – Liste des valeurs statistiques supportées par SumRep

# Tableau Comparatif des Outils

| Outil | Objectif | Entrée | Sortie | Langage /Mode d'utilisation |
|---|---|---|---|---|
| RDI | Quantifier les différences et établir les distances entre plusieurs répertoires de gènes | Vecteur contenant le nom des gènes et un vecteur (ou une matrice) contenant les annotations de chaque gène (donneur, type de cellule, etc) | — Matrice de distance entre chaque répertoire<br>— Diagramme de clusters | Package R |
| DivE | Prédiction du nombre d'espèces en fonction du nombre d'individus. Résoudre le problème des espèces non-observées dans un échantillon d'une population. | — Liste de modèles d'évolution<br>— Echantillon de la population étudiée sous forme de matrice à 2 colonnes : (espèce, nb individus)<br>— Nombre d'individus de la population totale | Estimation du nombre d'espèces/classes dans la population totale | Package R |
| ReCoLD | Etudier la diversité au sein d'un répertoire de séquences de récepteurs de cellules T, en identifiant les sous-séquences qui contribuent le plus aux différences. | Répertoire de séquences de récepteurs de lymphocytes T | — Matrice de dissimilarité<br>— Matrice JSD/de distance<br>— diagramme de clusters | Python |
| Sumrep | Donne un résumé détaillé d'un répertoire immunitaire en donnant plusieurs mesures | Répertoire immunitaire sous format AIRR | Plusieurs mesures de diversité (voir Figure 1) | Package R |
| ImmuneRef | Etudier la diversité d'un répertoire immunitaire en donnant plusieurs mesures interprétables | Répertoire immunitaire sous format AIRR | Plusieurs mesures de diversité | Package R |

# Tableau Comparatif des Outils

| | RDI | DivE | ReCoLD | SumRep | ImmuneRef |
|---|---|---|---|---|---|
| Matrice de distance | × | | | | |
| Détection de clusters/Diagramme de clusters | × | | × | | × |
| Indice de Shannon/Indice de Hill | | | | × | × |
| Prédiction d'évolution | | × | | × | |
| Génération de données ou modèles | × | | | × | |
| Jensen-Shannon Diversity (JSD) | | | × | × | × |
| Difficilté d'utilisation | ⋆ | ⋆ | ⋆⋆⋆ | ⋆⋆⋆ | ⋆⋆ |

# Compatibilité des données

Les outils cités plus haut ne prennent pas tous en entrée le même type de données.
En effet, par exemple, ImmuneRef et SumRep prennent en entrée des base de donnée en format AIRR cependant certaines annotations sont spécifiquement requises pour que l'outil reconnaisse les données. Il faut donc s'assurer de renommer les colonnes des dataframe et vérifier que les colonnes requises soient bien présentes.

## ImmuneRef

On peut vérifier la compatibilité avec une fonction disponible : compatibility_check.
Base de données en format AIRR avec le nom des colonnes requises :

- sequence : séquence en nucléotide
- sequence_aa : séquence en acides aminés
- junction : séquence CDR3 en nucléotide
- junction_aa : séquence CDR3 en acides aminés
- freqs : fréquence de chaque séquence dans la base de données
- v_call : annotation du gène V
- d_call : annotation du gène D
- j_call : annotation du gène J

## SumRep

Base de données en format AIRR avec le nom des colonnes requises :

- sequence : séquence en nucléotide
- junction : séquence CDR3 en nucléotide
- junction_aa : séquence CDR3 en acides aminés
- v_call : annotation du gène V
- d_call : annotation du gène D
- j_call : annotation du gène J
- stop_codon : si TRUE, la séquence n'est pas prise en compte
- vj_in_frame : si FALSE, la séquence n'est pas prise en compte

## RDI

Deux entrées :

- genes : vecteur contenant le nom des gènes
- seqAnnot : vecteur ou matrice contenant les annotations de chaque gene, sous format dataframe

## DivE

Deux entrées :

- Nom des gènes/espèces et nombre d'individus : liste contenant le nom de chaque espèce et le nombre d'individu pour chaque espèce
- Modèles : liste contenant les modèles à appliquer pour les prédictions (fournie dans le package).

# Tests et Résultats

Nous avons à notre disposition 3 répertoires immunitaires différents :

— Monoclonal : Répertoire d'une personne malade

— Oligoclonal : Répertoire d'une personne en début de maladie

— Polyclonal : Répertoire d'une personne saine

Il faut adapter le format des répertoires en fonction de l'outil utilisé. On peut effectuer ces modifications en manipulant les data frame directement dans le logiciel R.

## SumRep

### Pairwise distance distribution

On peut visualiser la distribution des distances de Levenshtein par paires de séquences pour chaque répertoire.

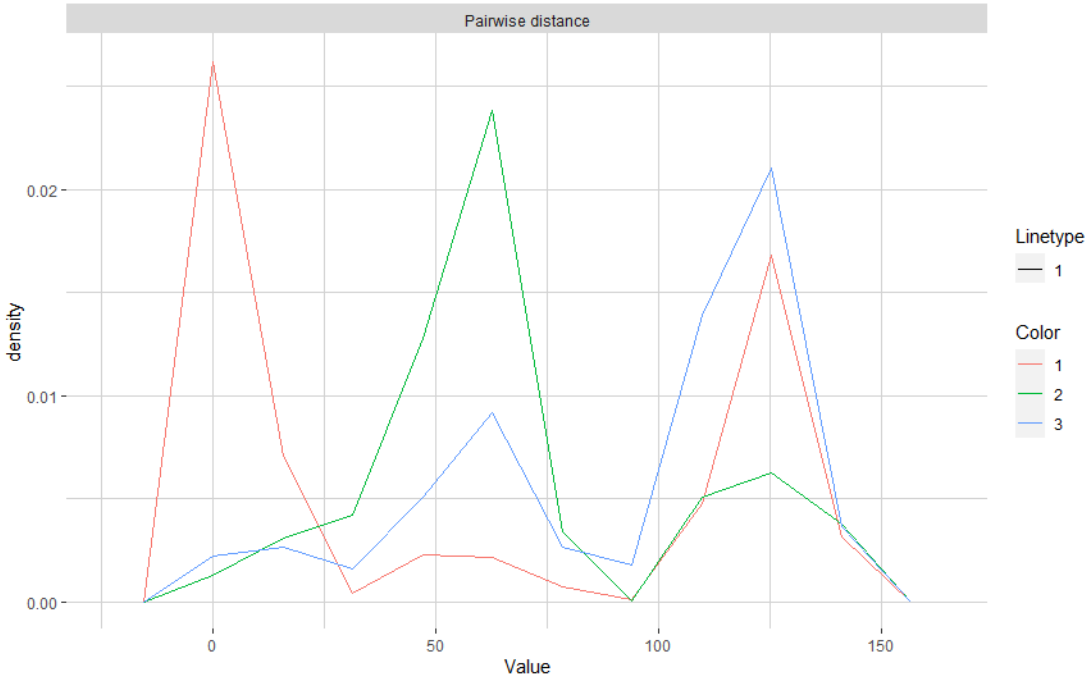


FIGURE 2 – Distribution des distances par paires de séquences. Le répertoire monoclonal est représenté en rouge, le répertoire polyclonal en bleu et le répertoire oligoclonal en vert.

Sur la Figure 2, on remarque que pour le répertoire monoclonal, plusieurs paires ont une distance de Levenshtein égale à zéro, ce qui montre que plusieurs séquences sont identiques et donc que la diversité est basse. Pour les deux autres répertoires, on observe des pics à environ 50 et 125, ce qui montre que la plupart des séquences ne sont pas identiques.

### Jensen-Shannon Divergence (JSD)

La divergence de Jensen-Shannon mesure la similarité entre les distributions de distances calculées précédemment. Plus cette valeur est élevée, plus les répertoires sont différents. A l'inverse, une valeur basse indique une faible similarité.

| | V1 | V2 | V3 |
|---|---|---|---|
| 1 | 0.0182679 | 0.18723982 | 0.29927822 |
| 2 | 0.1715475 | 0.01558668 | 0.12187294 |
| 3 | 0.3047391 | 0.14949334 | 0.01531556 |

FIGURE 3 – Matrice contenant les valeurs de JSD pour chaque paires de répertoires. Les lignes et colonnes 1, 2 et 3 correspondent respectivement aux répertoires monoclonal, oligoclonal et polyclonal

### Full Comparison

On peut effectuer une comparaison complète des caractéristiques de deux rpertoires. En effet, à l'aide de la fonction *compareRepertoires*(), SumRep applique toutes les comparaisons qu'il peut effectuer.

| | Comparison | Divergence |
|---|---|---|
| 1 | comparePairwiseDistanceDistributions | 0.34535050 |
| 2 | compareCDR3PairwiseDistanceDistributions | 0.24566214 |
| 3 | compareGCContentDistributions | 0.29682484 |
| 4 | KideraFactor1Divergence | 0.46945301 |
| 5 | KideraFactor2Divergence | 0.39054330 |
| 6 | KideraFactor3Divergence | 0.42298817 |
| 7 | KideraFactor4Divergence | 0.23059651 |
| 8 | KideraFactor5Divergence | 0.25610638 |
| 9 | KideraFactor6Divergence | 0.40815970 |
| 10 | KideraFactor7Divergence | 0.32880331 |
| 11 | KideraFactor8Divergence | 0.45796563 |
| 12 | KideraFactor9Divergence | 0.28885867 |
| 13 | KideraFactor10Divergence | 0.25023692 |
| 14 | AtchleyFactor1Divergence | 0.11709551 |
| 15 | AtchleyFactor2Divergence | 0.15422454 |
| 16 | AtchleyFactor3Divergence | 0.14430094 |
| 17 | AtchleyFactor4Divergence | 0.15363639 |
| 18 | AtchleyFactor5Divergence | 0.15490659 |
| 19 | compareAliphaticIndexDistributions | 0.29683030 |
| 20 | compareGRAVYDistributions | 0.45267629 |
| 21 | comparePolarityDistributions | NA |
| 22 | compareChargeDistributions | NA |
| 23 | compareBasicityDistributions | NA |
| 24 | compareAcidityDistributions | NA |
| 25 | compareAromaticityDistributions | NA |
| 26 | compareBulkinessDistributions | NA |
| 27 | compareInFramePercentages | 0.00000000 |
| 28 | compareAminoAcidDistributions | 0.92327129 |
| 29 | compareAminoAcid2merDistributions | 1.48600474 |
| 30 | compareCDR3LengthDistributions | 0.35850290 |
| 31 | compareVGeneDistributions | 1.95670996 |
| 32 | compareJGeneDistributions | 1.75516512 |
| 33 | compareVGene3PrimeDeletionLengthDistributions | NA |
| 34 | compareJGene5PrimeDeletionLengthDistributions | NA |
| 35 | compareDGeneDistributions | 1.69922621 |
| 36 | compareVDJDistributions | 2.00000000 |
| 37 | compareDGene3PrimeDeletionLengthDistributions | NA |

FIGURE 4 – Comparaison complète entre le répertoire monoclonal et le répertoire polyclonal

## ImmuneRef

## DivE

Pour utiliser cet outil, on ne prend en compte que les séquences et le nombre d'occurrence de chaque séquence. L'outil DivE renvoie une estimation du nombre d'espèces pour une population d'une taille 100 fois plus grande que la taille de l'échantillon.

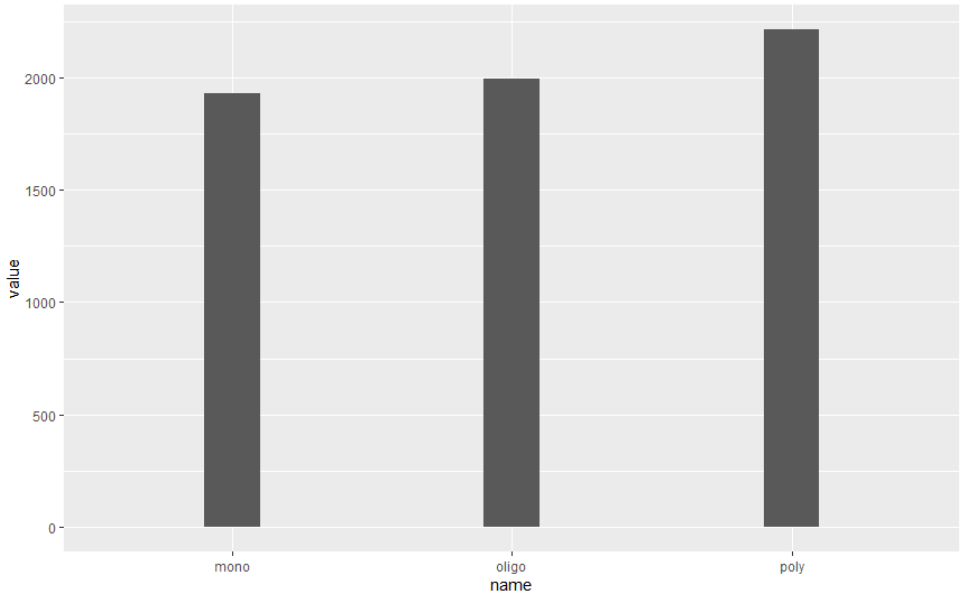


FIGURE 5 – Estimation du nombre d'espèces en fonction du répertoire

On remarque que l'estimation est la plus basse pour le répertoire monoclonal, plus grande pour le répertoire polyclonal et intermédiaire pour le répertoire oligoclonal, ce qui correspond aux résultats attendus.