Analyse de données génomiques : orientation des gènes répétés en tandem et distances intergéniques

Julien Stihle

UNIVERSITÉ DE STRASBOURG

gm Génétique Moléculaire = Génomique = Microbiologie

Rapport de stage

Julien STIHLE

Analyse statistique de données génomiques: orientation des gènes répétés en tandem et distances intergéniques

Institut de Botanique

Période du 21 juin au 19 août 2011

Maitre de stage : Laurence DESPONS

Tuteur : Armelle GUILLOU

REMERCIEMENTS

Mes remerciements s'adressent en premier lieu à mon maître de stage, Madame Laurence DESPONS, enseignant-chercheur à l'Université de Strasbourg, pour sa confiance et ses conseils qui m'ont permis de progresser sans cesse durant ces 2 mois de stage. Je remercie également mon tuteur pédagogique, Monsieur Nicolas POULIN, pour ses nombreux conseils en statistique.

Ce stage a nécessité tout au long de sa durée l'aide et le soutien de Véronique LEH LOUIS et Zlatyo UZUNOV.

Je tiens à exprimer toute ma reconnaissance à Monsieur Serge POTIER, directeur de l'UMR 7156, pour son accueil au sein de l'équipe et sa disponibilité.

J'exprime également ma gratitude à l'égard de l'ensemble du personnel pour leur précieuse aide ainsi que leur sympathie qui ont favorisées mon intégration dans le laboratoire.

Sommaire

I. Présentation du laboratoire d'accueil

J'ai effectué mon stage de deux mois, de la période du 21 juin au 19 août 2011, au sein de l'Institut de Botanique de Strasbourg, situé 28 rue Goethe. Dans le cadre de mon stage, j'ai été accueilli dans une équipe très dynamique, dirigé par Mr Serge Potier, à l'unité mixte de recherche 7156 (UMR7156) de Génétique Moléculaire, Génomique, Microbiologique (GMGM). Cette UMR est divisée en deux départements, avec un effectif total de 60 personnes réparties dans six équipes. Le département "Génétique moléculaire et cellulaire" est situé à l'Institut de Physiologie et de Chimie Biologie (IPCB), au 21 rue Descartes à Strasbourg et est constitué trois équipes. Le second département "Micro-organismes, génomes, environnement", est situé dans l'enceinte de l'Institut de Botanique. Cette UMR a pour objectif l'étude des mécanismes fondamentaux des fonctionnements et dysfonctionnements des génomes de cellules procaryotiques et eucaryotiques, ainsi que de divers processus intracellulaires.

Tout au long de mon stage, j'ai été amené à travailler dans l'équipe du Pr Jean-Luc Souciet. Cette équipe s'intéresse à l'évolution des génomes de micro-organismes et plus précisément à l'étude des mécanismes conduisant à la formation et à la disparition des gènes, ainsi que ceux impliqués dans le remodelage des génomes. Mon étude statistique a été orientée dans le domaine de la génomique par Mme Laurence Despons, maître de stage, Maître de Conférences à l'Université de Strasbourg.

II. Présentation du sujet

Dans le cadre de mon stage, j'ai été amené à réaliser une analyse statistique sur des données biologiques. Pour travailler sur ce type de données, il a tout d'abord fallu comprendre le sujet de l'étude. L'étude relevait du domaine de la génomique (l'étude des génomes) et portait sur les gènes répétés en tandem présents dans les génomes de 11 espèces de levures. Je vais, dans un premier temps, définir les notions nécessaires à la compréhension du sujet, puis expliquer l'intérêt des recherches et l'étude statistique que j'ai réalisée.

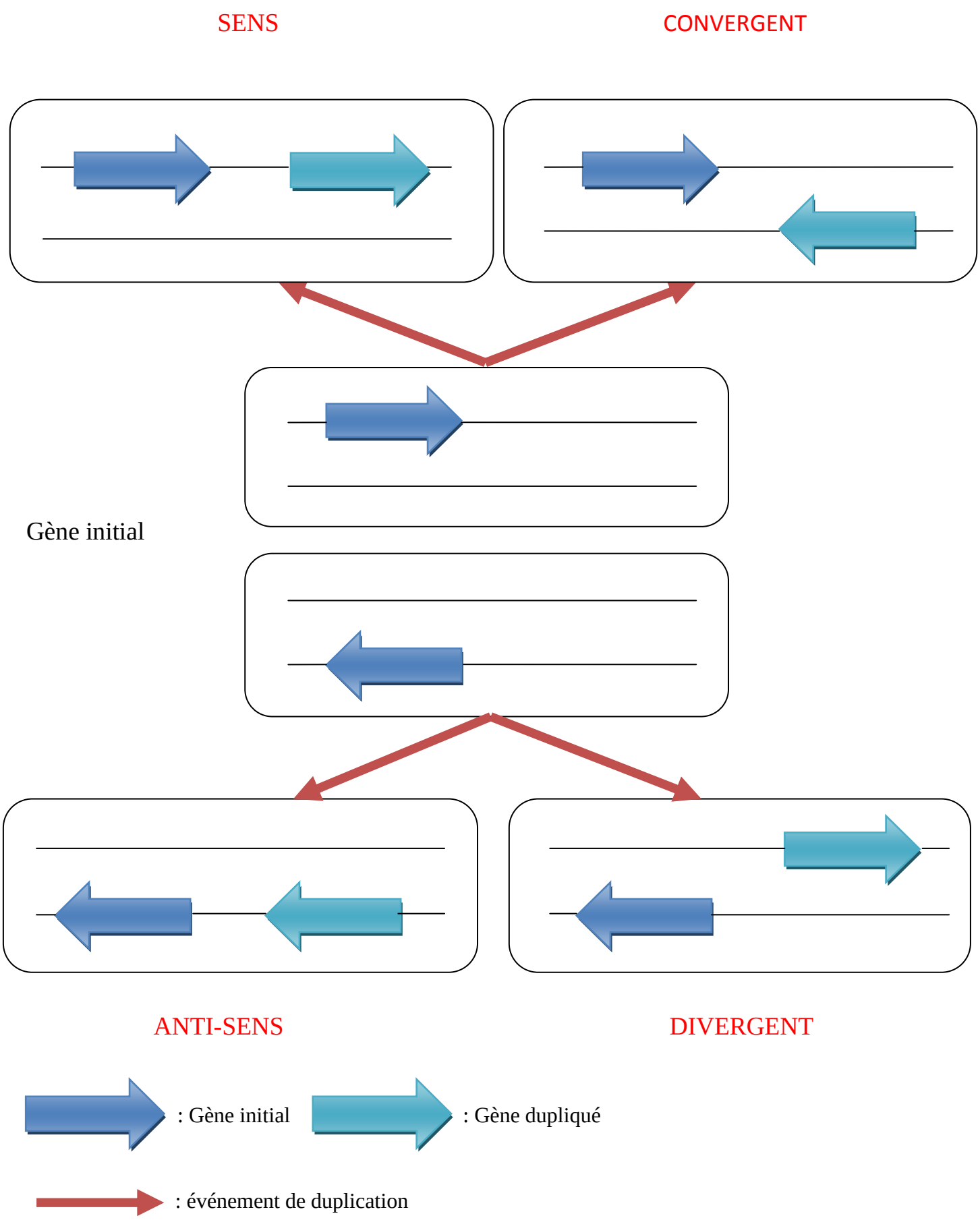
Chaque espèce de levure étudiée possède un génome constitué de plusieurs chromosomes. Chaque chromosome est formé de deux brins d'ADN sur lesquels sont situés les gènes. Au cours du temps, les génomes évoluent par différents mécanismes. Le mécanisme d'évolution qui nous intéresse ici est la duplication de gène. Ce mécanisme permet de copier la séquence d'un gène et d'insérer la copie n'importe où dans le génome. Si l'insertion se fait sur le même chromosome et de surcroît juste à côté du gène initial, on parle alors de duplication de gène en tandem, et la structure formée est appelée « tandem de gènes » (ou « tandem »). On obtient ainsi un couple de gènes qui possèdent une homologie de séquence (gènes qui « se ressemblent » en séquence car dérivant d'un même gène ancêtre). Le tandem peut contenir plus de deux éléments si le phénomène de duplication se produit à plusieurs reprises. Deux copies de gènes peuvent avoir plusieurs orientations l'une par rapport à l'autre (voir Figure 1). Lorsque les copies sont sur le même brin d'ADN (en sens ou anti-sens), l'orientation est dite « directe ». Lorsqu'elles sont sur les deux brins (convergentes ou divergentes), l'orientation est « inversée ».

Tous les génomes actuellement connus possèdent des gènes dupliqués car leur mécanisme de formation, la duplication, est un moteur de l'évolution (Ohno, 1970[1]). En effet, une des copies dupliquées peut conserver la fonction du gène ancestral, tandis que l'autre peut évoluer et acquérir une nouvelle fonction. Dans le cas de la seconde copie, nous sommes face à la « naissance » d'un nouveau gène. Les gènes dupliqués en tandem, bien qu'assez fréquents dans les génomes des vertébrés (jusqu'à plus de 20% des gènes totaux ; Pan et Zhang, 2008[2]), ont été peu étudiés jusqu'à présent. De ce fait, leur mécanisme de formation n'est pas encore connu. L'objectif de l'étude menée dans mon laboratoire d'accueil est donc de comprendre le ou les mécanisme(s) moléculaire(s) de la duplication de gène en tandem. L'approche utilisée est la génomique comparative et fait appel essentiellement à des outils informatiques (analyse *in silico*). Ainsi, les génomes de 11 espèces différentes de levures ont été analysés pour rechercher les tandems de gènes (Despons *et al.*, 2010[3] et Despons *et al.*, 2011[4]). L'analyse informatique du catalogue des tandems obtenu, qui est en cours, devrait permettre de mettre en évidence des points communs entre ces tandems (motifs, taille…) pour proposer des hypothèses sur les mécanismes de formation des tandems.

Je me suis intéressé, d'un point de vue statistique, aux données d'orientation des gènes et de distance entre les gènes. Les résultats d'une comparaison entre les gènes localisés dans des tandems et hors des tandems tendent à montrer que : (i) l'orientation directe serait favorisée pour les gènes en tandem et (ii) la distance moyenne entre deux gènes (distance intergénique) serait plus élevée dans les tandems. Si ces deux hypothèses sont validées par l'analyse statistique, le mécanisme de duplication en tandem produirait plus fréquemment des paires de gènes sur le même brin d'ADN et éloignerait les gènes.

[1]Ohno S. (1970) Evolution by gene duplication. Springer-Verlag, Berlin-Heidelberg-New York
[2]Pan D., Zhang L. (2008) Tandemly arrayed genes in vertebrate genomes. Comp. Funct. Genomics 545269
[3]Despons L., Baret PV., Frangeul L., Louis VL., Durrens P., Souciet JL. (2010) Genome-wide computational prediction of tandem gene arrays: application in yeasts. BMC Genomics 11:56
[4]Despons L., Uzunov Z., Louis VL. (2011) Tandem gene arrays, plastic chromosomal organizations. CR Biol. 334:639

Figure 1. Orientation des gènes en tandem après une duplication d'un gène initial



III. Présentation des données

Le but de l'étude statistique réalisée est de déterminer si il existe une corrélation entre différentes caractéristiques d'un couple de gènes comme la distance intergénique, l'orientation des gènes, le fait d'être en tandem ou encore l'espèce.

J'avais à ma disposition un tableau comprenant 10 colonnes (Table1) qui contient toutes les caractéristiques de positions : le numéro de l'espèce (Species), le numéro du chromosome où est situé le couple de gène (Chromosome), le nom du premier gène (CDS_1), le nom du second gène (CDS_2), la coordonnée de début de la distance intergénique (Start), la coordonnée de fin de la distance intergénique (End), la valeur de la distance intergénique (Distance), une réponse binaire pour le fait d'être en tandem ou non (InTGA=1 indique que les gènes CDS_1 et CDS_2 sont en tandem, InTGA=0 indique que ces mêmes gènes ne sont pas en tandem), l'orientation entre les deux gènes (Orientation) et le numéro du tandem si les gènes sont en tandem (TGAid). La taille du tableau de données est assez conséquente : 10 colonnes et 66087 lignes (qui représentent tous les intervalles de gène identifiés dans les 11 espèces). J'ai déterminé les colonnes à prendre en considération lors de l'étude, d'une part pour effectuer une vérification des données ou des tests statistiques et d'autre part pour avoir des repères dans notre tableau. Etant donné les questions scientifiques posées, il est évident que les colonnes importantes à la réalisation de l'étude sont l'espèce, la distance intergénique, l'orientation du couple de gène et la colonne identifiant les gènes en tandems (InTGA).

Pour simplifier la programmation j'ai décidé de changer le nom de l'espèce par des numéros, on a respectivement de 1 à 11 les espèces de levures suivantes : Arad, Cagl, Deha, Ergo, Klla, Klth, Piso, Sace, Sakl, Yali et Zyro. De même les orientations sont notées de 1 à 4 respectivement pour l'orientation en sens, en anti-sens, convergente et divergente. Voici un extrait du tableau que l'on utilisera:

Table1 : Extrait du tableau de données

Species	Chromosome	CDS_1	CDS_2	Start	End	Distance	Orientation	InTGA	TGAid
1	1	ARAD0A00110g	ARAD0A00132g	1268	1779	511	2	0	
1	1	ARAD0A00132g	ARAD0A00154g	2441	3702	1261	4	0	
1	1	ARAD0A00154g	ARAD0A00176g	4043	4578	535	1	0	
1	1	ARAD0A00176g	ARAD0A00198g	5078	7814	2736	1	0	
1	1	ARAD0A00198g	ARAD0A00220g	8302	10017	1715	1	0	
1	1	ARAD0A00220g	ARAD0A00242g	10292	11003	711	1	0	
1	1	ARAD0A00242g	ARAD0A00264g	11479	13504	2025	3	0	545/546
1	1	ARAD0A00264g	ARAD0A00286g	15510	16931	1421	2	1	546
1	1	ARAD0A00286g	ARAD0A00308g	18931	20839	1908	2	0	
1	1	ARAD0A00308g	ARAD0A00330g	21624	22285	661	4	0	
1	1	ARAD0A00330g	ARAD0A00352g	22782	23054	272	1	0	
1	1	ARAD0A00352g	ARAD0A00374g	25240	25356	116	3	0	
1	1	ARAD0A00374g	ARAD0A00396g	25948	26455	507	4	0	
1	1	ARAD0A00396g	ARAD0A00418g	27072	27494	422	3	0	

Nous observons dans le tableau (Table1) une ligne en rouge qui correspond à une distance intergénique mesurée sur un tandem (InTGA=1). La ligne représenté en bleu fait référence à une distance intergénique mesurée sur un couple de gène hors tandem (InTGA=0). Après avoir étudié le tableau, je me suis rendu compte de quelques incohérences. Par exemple, la base de données utilisée pour identifier les tandems avait repéré 622 tandems dans les génomes des 11 espèces de levures, mais dans le tableau reçu certains tandems n'apparaissaient pas. Il a donc fallu identifier les éléments manquants puis comprendre pour quelle(s) raison(s) ils n'apparaissaient pas. Pour cela je me suis aidé de la colonne TGAid qui m'a permis d'identifier chaque élément manquant. Puis j'ai utilisé une base de données contenant tous les gènes par espèce pour identifier l'erreur commise dans la colonne InTGA.

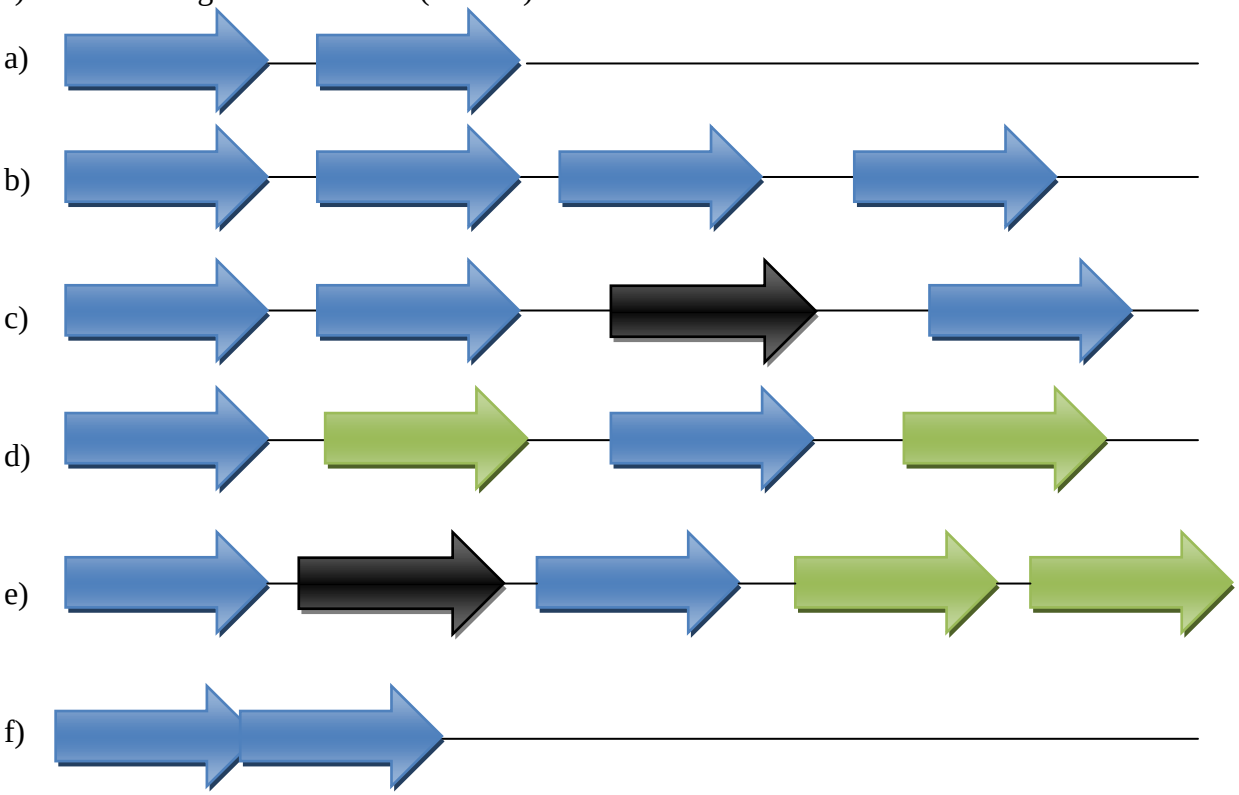
Je me suis également rendu compte que dans la notion de gène en tandem plusieurs cas étaient possibles. Il a fallu définir de façon très rigoureuse la notion de gène en tandem. Pour cela je me suis basé sur la définition et les propriétés que Mme Despons m'avait données :

1) Un tandem est un couple de gènes consécutifs où les gènes possèdent une homologie (fig2a)
2) Il est possible que les tandems soient formés de plus de deux éléments si l'homologie se poursuit avec le gène suivant (fig2b)
3) Il existe des tandems particuliers ayant des gènes intercalés entre deux gènes homologues (fig2c). Cette situation ne se produit pas très souvent.
4) Il est également possible d'avoir des tandems intercalés dans des tandems (fig2d). Comme on peut le voir dans le tableau précédent, nous observons une ligne avec « 545/546 » dans la colonne TGAid. Cela signifie qu'il existe un enchevêtrement de tandem (fig2d) ou bien deux tandems contigus (fig2e). L'ordre des gènes est alors le suivant : un gène du tandem 545, un gène intercalé (qui n'appartient à aucun tandem), un autre du 545, un gène du tandem 546 et enfin un dernier du 546. Dans le tableau initial la colonne InTGA contient un « 1 » à la ligne « 545/546 » alors qu'il ne s'agit pas d'une distance entre gènes d'un même tandem.
5) Il est de plus possible d'observer des superpositions de gènes (fig2f)

Pour éviter de reproduire l'erreur observée en 4), Mme Despons a décidé de travailler uniquement sur les tandems n'ayant pas d'éléments intercalés (cas c et d éliminés). J'ai donc modifié le tableau pour obtenir une nouvelle colonne InTGA correspondant à notre définition de tandem de gènes : il s'agit de gènes côte à côte qui possèdent une homologie au niveau de leur séquence. Sur un total de 622 tandems, 31 ont donc été renommés en « hors tandem », soit 5% des données de départ. De plus les couples de gènes superposés ont été éliminés de l'étude comme cela avait déjà été réalisé dans une étude auparavant. J'ai remarqué que ce cas ne se produisait que très peu et de plus que ce phénomène n'intervenais pas sur les gènes en tandem avec nos données.

Figure 2 : Exemples de structures en tandem observée dans les génomes

Lorsque les deux gènes sont en tandem, ils sont représentés de la même couleur a) et b) Cas classiques de tandem de gènes avec une orientation en sens. c) Représentation d'un tandem interrompu (en sens) d) Deux tandems sont intercalés l'un dans l'autre e) Deux tandems contigus avec un gène intercalé dans un tandem f) Cas de deux gènes en tandem (en sens) se chevauchant



IV. Analyse préliminaire des données

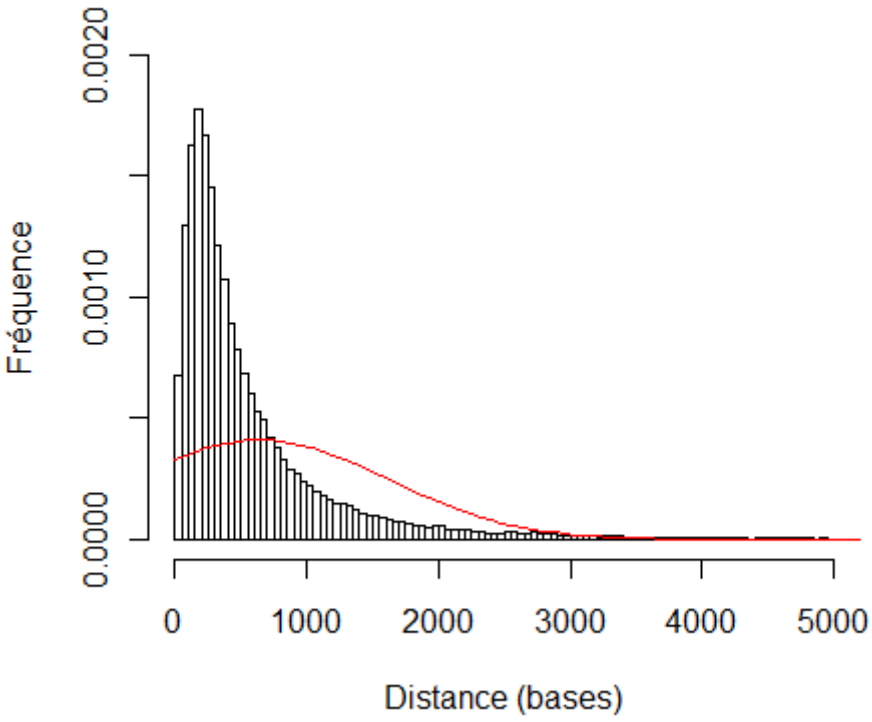
La réalisation de toutes mes analyses statistiques ont été réalisées à l'aide du logiciel R avec un risque de première espèce de 5%. Une des questions posées est de savoir si l'orientation des gènes avait un effet significatif sur les distances intergéniques. Au début d'une telle étude statistique, il est important de connaître la distribution de nos données et plus particulièrement de savoir si elles suivent une loi normale. Cette information est essentielle pour déterminer la nature des tests à utiliser : tests paramétriques (dans le cas d'une loi normale et homoscédasticité) ou tests non paramétriques.

A. Normalité des données

1. Méthodes graphiques

Au cours de l'étude, il est possible d'utiliser des représentations graphiques de nos données ceci pour avoir une meilleure approche de la distribution de nos données, particulièrement quand l'effectif est très élevé. Un graphique ne répond jamais à un problème, mais il peut permettre, d'anticiper certaines réponses et par la même occasion de découvrir certaines anomalies qui peuvent donner lieu à de nouvelles problématiques. Le logiciel R nous permet d'utiliser une représentation graphique des données sous forme d'un histogramme. Nous observons ainsi la distribution de ces données. Elle peut être comparée à une loi normale en superposant sa distribution avec une loi normale de moyenne **μ** et variance **σ**, où **μ** représente la moyenne des distances intergéniques, et **σ** la variance des distances intergéniques. L'affichage de l'histogramme des distances se fait à l'aide de la commande **hist()**. Ensuite la commande **plot()** avec l'option **add=TRUE** superpose notre loi normale à notre histogramme. Nous en déduirons que, plus la courbe est proche de la cime de l'histogramme et plus nous aurons une compatibilité de nos données avec une distribution normale.

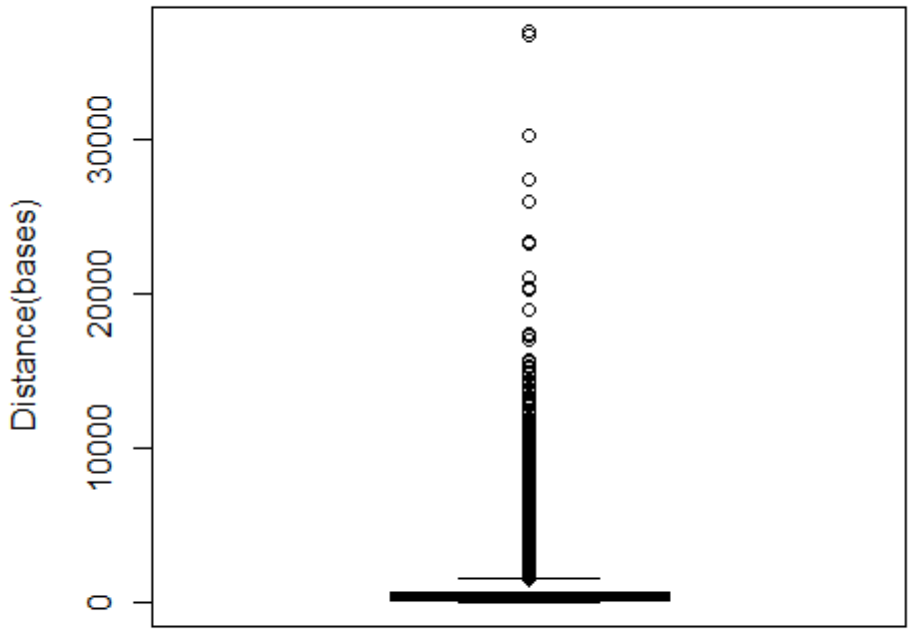
Figure 3 : Histogramme des distances



D’après l’histogramme des distances intergéniques (Figure 3) nous remarquons que la courbe de la loi normale n’est pas superposée avec la distribution des distances. Nous supposons donc que les données ne sont pas distribuées selon une loi normale. Cependant il existe plusieurs autres méthodes graphiques pour confirmer cette hypothèse de non appartenance à une loi normale.

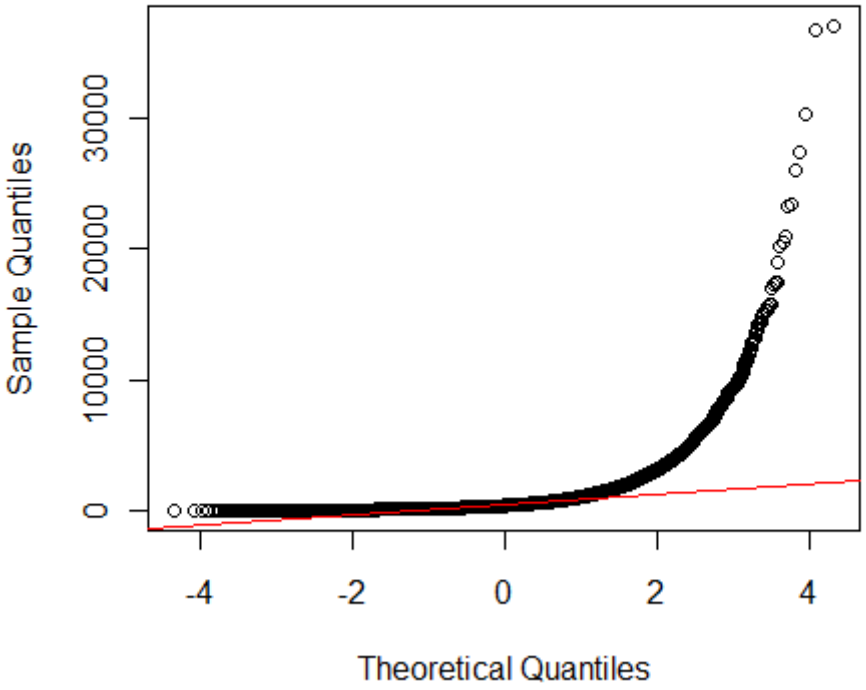
Il est également possible d’utiliser une boîte à moustache des distances (Figure 4) pour repérer les anomalies de distribution. On utilise pour cela la commande **boxplot()**. Notre boîte à moustache doit être symétrique dans le cas de la loi normale. De plus, il ne doit pas y avoir trop de valeurs très éloignées des moustaches.

Figure 4 : Boite à moustaches de nos distances



Nous observons que la boîte à moustache des distances n’est pas symétrique. Cette non-symétrie des données, va à l’encontre d’une hypothèse de normalité des données. Il existe une troisième représentation graphique qui est plus représentative de la normalité des données : le graphique des quantiles (Figure 5). Il s’agit de comparer notre échantillon observé avec la distribution théorique.

Figure 5 : Graphique des quantiles



D'après le graphique des quantiles et la droite de Henry (en rouge), nous remarquons que les problèmes résident sur les queues de distributions. Après avoir réalisé une première approche graphique sur la normalité des données, je vais analyser les données des distances à l'aide de tests statistiques pour avoir une approche plus précise.

2. Test statistique de normalité

Pour mettre en évidence une normalité des données, il existe plusieurs tests statistiques : le test de Shapiro-Wilk, le test de Lilliefors, le test d'Anderson-Darling, le test de D'Agostino ou encore le test de Jarque-Bera. Parmi tous ces tests le plus fréquemment utilisé est le test de Shapiro-Wilk pour sa puissance avec des échantillons qui ont peu d'observations. Les autres tests sont soit des variantes plus puissantes du test de Kolmogorov-Smirnov (Lilliefors et Anderson-Darling), soit basés sur les coefficients d'asymétrie et d'aplatissement (D'Agostino et Jarque-Bera).
Dans notre cas, au vu du nombre de données (65762), le test de Shapiro-Wilk n'est pas approprié car il est efficace uniquement pour un échantillon de 3 à 5000 observations. J'ai donc décidé d'utiliser un **test de D'Agostino** (**dagoTest()**) car les tests basés sur la statistique du test de normalité de Kolmogorov-Smirnov (test de Lilliefors et test d'Anderson-Darling) sont sensibles aux ex-æquo dans l'échantillon. Etant donné que mes valeurs sont des nombres entiers, il n'est pas rare d'observer des valeurs répétées. Dans notre cas nous utiliserons plutôt le test de D'Agostino que le test de Jarque-Bera, de part sa puissance sur des échantillons plus grands.

3. Test de D'Agostino

Ce test est également connu sous le nom de test K2 (K-squared) de D'Agostino-Pearson. Le test compare les deux coefficients par rapport à 0. On conclut que la distribution ne suit pas une loi normale, si les deux coefficients diffèrent simultanément de la valeur 0.

On reproche souvent au test de D'Agostino qu'il ne permet pas directement de comprendre la nature de la déviation de la loi normale en cas de rejet de l'hypothèse nulle. Il est donc nécessaire de compléter l'analyse avec l'étude individuelle des coefficients pour connaître la nature de la déviation.
Le test consiste à centrer et réduire les deux coefficients (asymétrie et aplatissement) de manière à obtenir des valeurs z_1 et z_2 distribuées asymptotiquement selon une loi normale (0;1). Des corrections supplémentaires sont réalisées de manière à rendre l'approximation normale plus efficace.

L'hypothèse nulle H0 de ce test : distance $\rightsquigarrow \mathcal{N}$ (**μ**, **σ**)
L'échantillon est distribué selon une loi normale de moyenne **μ** et de variance **σ**.
L'hypothèse alternative H1 de ce test : distance $\not\rightsquigarrow \mathcal{N}$ (**μ**, **σ**)
L'échantillon n'est pas distribué selon une loi normale de moyenne **μ** et de variance **σ**.

Une première transformation est effectuée sur le coefficient d'asymétrie (g_1). Les calculs successifs sont les suivants :

$$g_1=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i-}\bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_{i-}\bar{x})^2)^{\frac{3}{2}}}$$

$$A=g_1\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}$$

$$B=\frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$C=\sqrt{2(B-1)}-1$$

$$D=\sqrt{C}$$

$$E=\frac{1}{\sqrt{\ln(D)}}$$

$$F=\frac{A}{\sqrt{\frac{2}{C-1}}}$$

$$z_1 = Eln(F+\sqrt{F^2+1})$$

Nous procédons de manière similaire pour le coefficient d'aplatissement.

$$g_2=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_i(\frac{x_i-\bar{x}}{s})^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

$$G=\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$H=\frac{(n-2)(n-3)g_2}{(n+1)(n-1)\sqrt{G}}$$

$$J=\frac{6(n^2-5n+2)}{(n+7)(n+5)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$K=6+\frac{8}{3}\left(\frac{2}{J}+\sqrt{1+\frac{4}{J^2}}\right)$$

$$L=\frac{1-\frac{2}{K}}{1+H\sqrt{\frac{2}{K-4}}}$$

$$z_2=\frac{\left(1-\frac{2}{9K}\right)-L^{\frac{1}{3}}}{\sqrt{\frac{2}{9K}}}$$

z_1 et z_2 suivent asymptotiquement une loi normale $N(0,1)$. La statistique du test est la combinaison :

$K2=z_1^2+z_2^2$

Elle suit asymptotiquement une loi du $\chi 2$ à 2 degrés de libertés. L'incompatibilité de la distribution évaluée avec la loi normale est d'autant plus marquée que la statistique $K2$ prend une valeur élevée. Pour un risque α, la région critique du test s'écrit :

R.C. : $K2 > \chi^2_{1-\alpha}(2)$

Pour α= 0:05, le seuil critique est $\chi^2_{0.95}(2)=5,99$.

Pour utiliser le test de D'Agostino avec le logiciel R, il faut tout d'abord charger la bibliothèque « fBasics ».

```
> dagoTest(Distance2)

Title:

 D'Agostino Normality Test

Test Results:

  STATISTIC:

    Z3  | Skewness: 254.6707

  P VALUE:

    Skewness Test: < 2.2e-16
```

Lorsque la p-valeur inférieur à 2.2*10^-16 l'hypothèse de normalité des distances avec le test de D'Agostino est rejetée.

Conclusion : Nous retrouvons bien le résultat observé sur le graphique des quantiles, c'est-à-dire, nous rejetons l'hypothèse de normalité des données. Dans ce cas, avant de poursuivre l'étude statistique sur ces données, il est intéressant d'essayer de normaliser ces données à l'aide de transformation. L'intérêt de telles transformations est l'utilisation de tests paramétriques qui sont plus puissants que les tests non-paramétriques.

4. Transformations des données

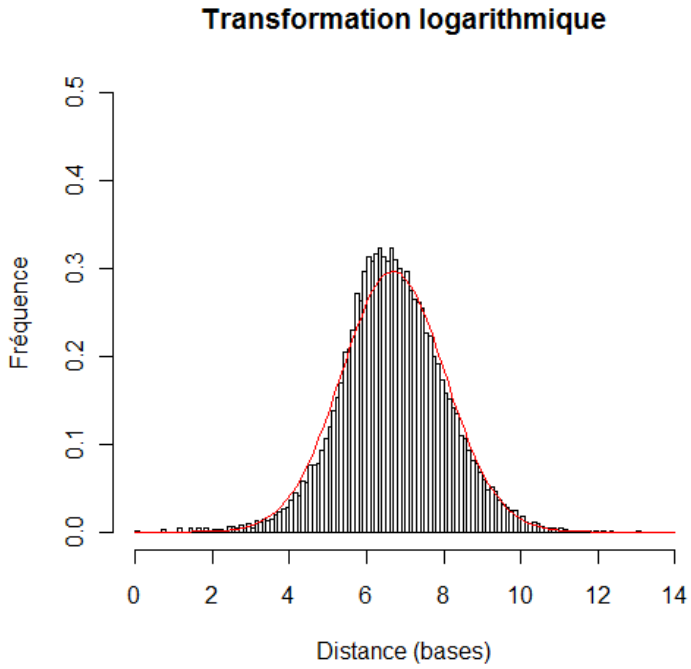
Il existe plusieurs transformations qui permettent de normaliser les données : la transformation logarithmique, la transformation racine ou encore la transformation de Box-Cox. Ces trois transformations nécessitent d'utiliser des variables à valeurs strictement positives. Etant donné que j'ai supprimé toutes les valeurs négatives de mes distances, il ne me reste que le cas des valeurs égales à 0. Pour supprimer ces cas, j'ai augmenté les valeurs des distances de 1. Ainsi la variance est inchangée et par la même occasion la moyenne ne se voit pas trop affectée (une unité de plus qu'avant).

a) Transformation logarithmique

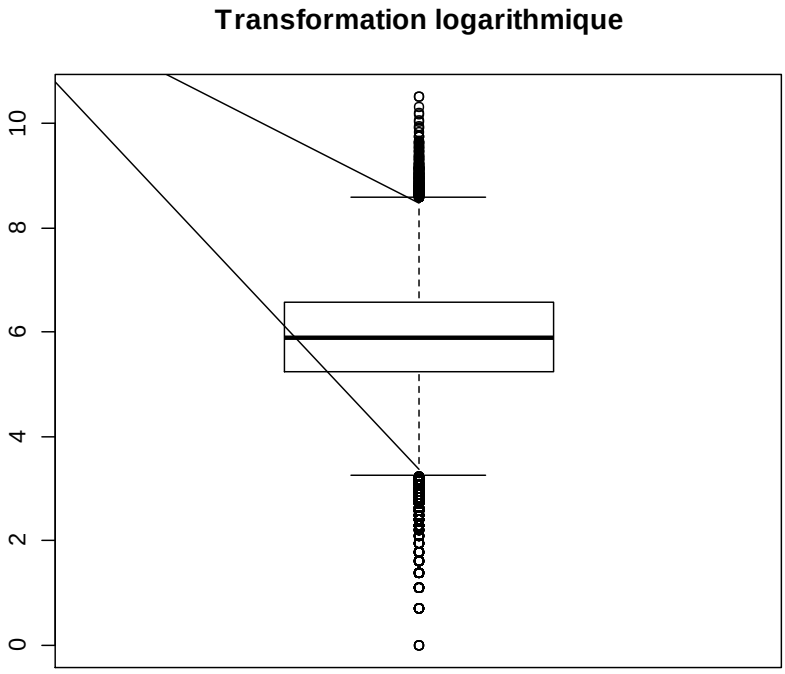
La transformation logarithmique normalise les données avec la transformation suivante :

$y' = \ln(y)$, car ici y est strictement positif.

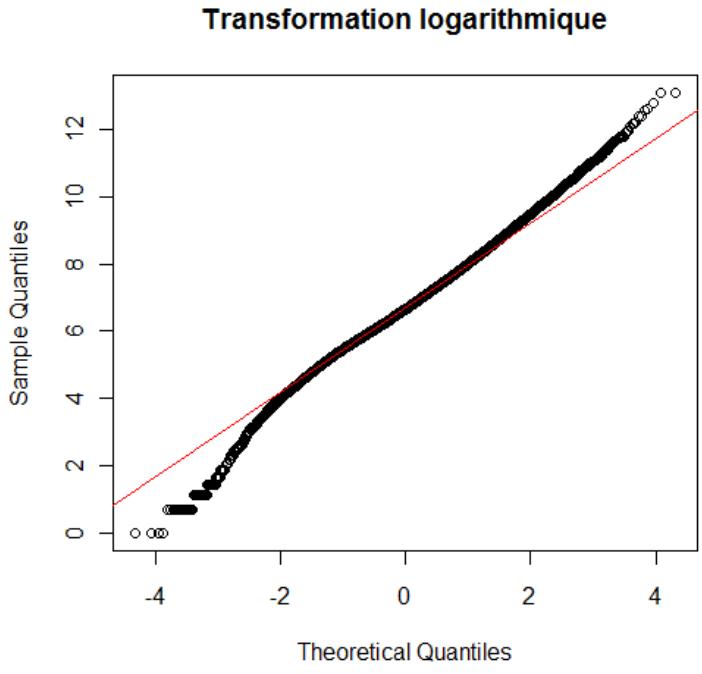
Après avoir obtenu les nouvelles données, il est maintenant question de savoir si elles ont été suffisamment normalisées. Nous recommençons notre analyse graphique et notre analyse statistique comme pour les données initiales afin d'observer les éventuelles améliorations qui ont été faites à l'aide de cette transformation. Les nouvelles représentations graphiques sont les suivantes :



Nous observons une amélioration assez nette vers une normalisation des données. En effet, les données sont très proches du comportement de la courbe de la loi normale (courbe rouge).



Une symétrisation de la boîte à moustacheest aussi observée.

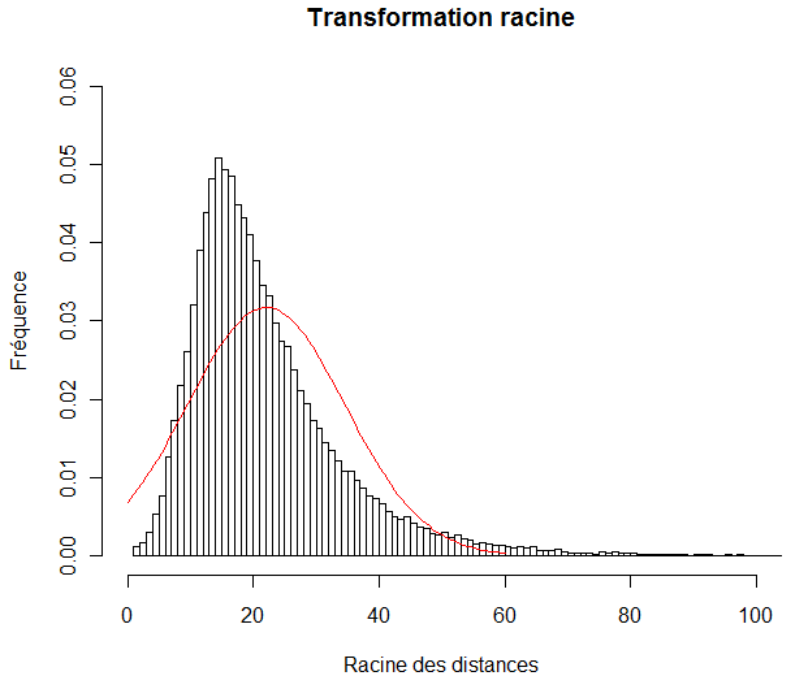


Nous remarquons cependant que le graphique des quantiles, n'est pas suffisamment convaincant. Il est donc nécessaire de faire un test statistique pour déterminer si ces données suivent réellement une loi normale. Pour les mêmes raisons que précédemment j'ai utilisé le test de D'Agostino. Le résultat est le même : la p-valeur est inférieure à 2.2*10^-16. Cela ne signifie pas qu'il n'y a eu aucune amélioration d'un point de vu statistique étant donné qu'il ne s'agit pas de valeurs précises mais plutôt que la transformation n'a pas été suffisamment normalisante.

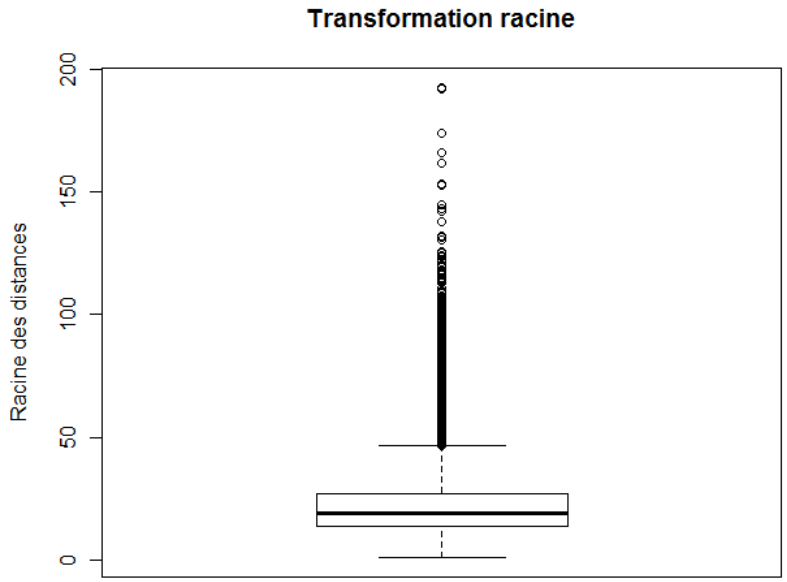
b) Transformation racine

Comme notre première tentative de transformation a échoué, nous avons utilisé une autre transformation : la transformation racine :

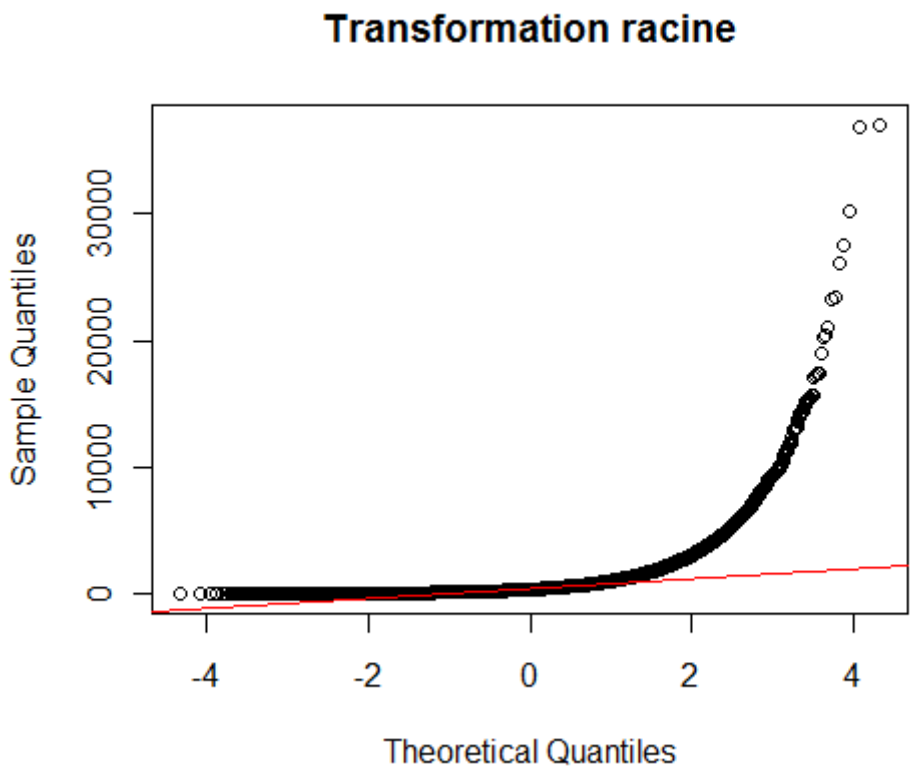
$y' = \sqrt{y}$, y étant toujours strictement positif.



Graphiquement le résultat obtenu avec la transformation racine est moins bon que celui obtenu avec la transformation logarithmique.



De même la boite à moustache est à nouveau asymétrique.



Le graphique des quantiles nous indique également la non-normalité des données.

L'idée de non-normalité peut se confirmer à l'aide d'un test de D'Agostino. La p-valeur est inférieure à 2.2*10^-16. On rejette donc à nouveau l'hypothèse de normalité.

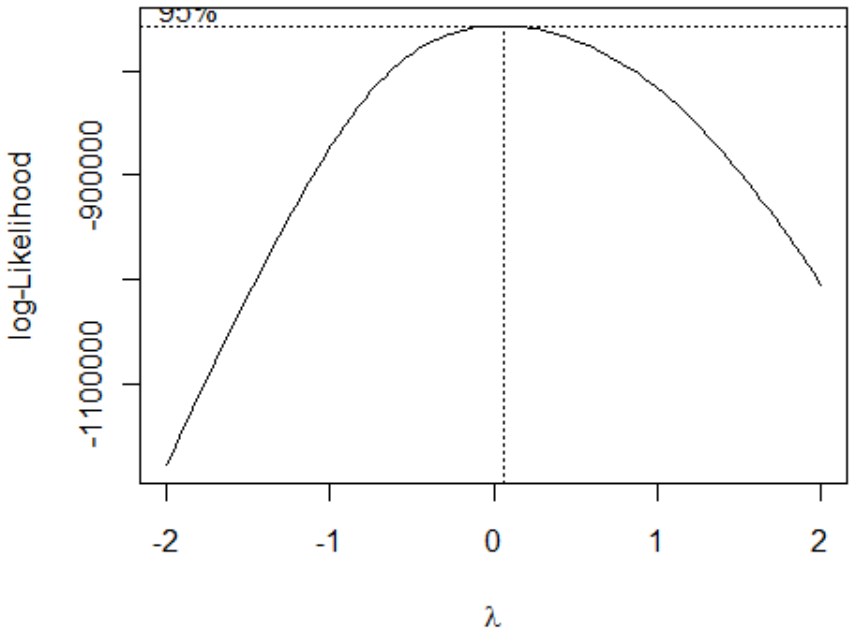
c) Transformation de Box-Cox

On transforme ces données à l'aide de la transformation de Box-Cox. Cette transformation est un mélange des deux transformations précédentes, et est définie comme suit :

$y' = \frac{(y^{\lambda} - 1)}{\lambda}$ si ($\lambda \neq 0$)

$y' = \ln(y)$ si ($\lambda = 0$)

La stratégie pour détecter facilement la valeur adéquate du paramètre λ dans la transformation est de balayer un grand nombre de valeurs de λ et de surveiller la valeur de r (log-vraisemblance) calculée sur la droite de Henry. On choisira la valeur λ^* qui maximise *r*. Pour obtenir une vue synthétique de la simulation, on construit généralement un graphique qui met en relation¸ λ (en abcisse) et *r* (en ordonnée) : il s'agit du Box-Cox Normality Plot.

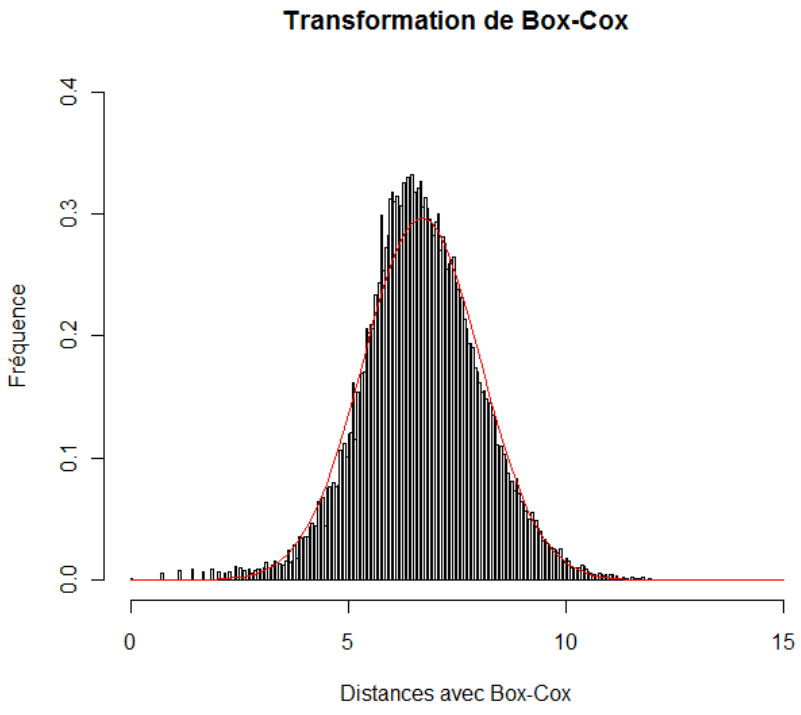


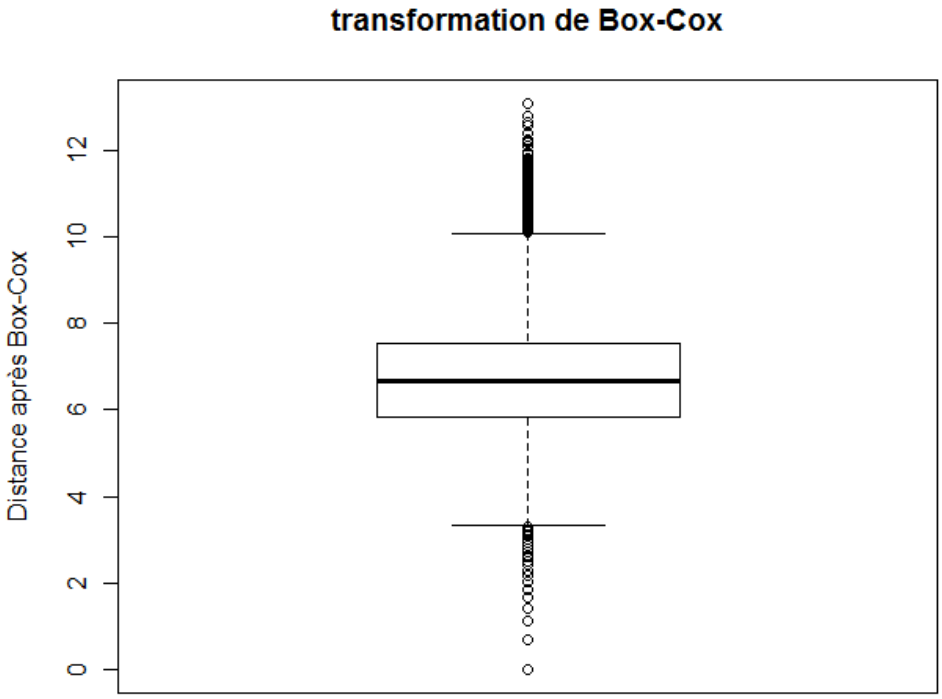
On obtient $\boldsymbol{\lambda} = 0.04$ si l'on observe les valeurs de $\boldsymbol{\lambda}$ sur [-2,2] avec un pas de 1/100.

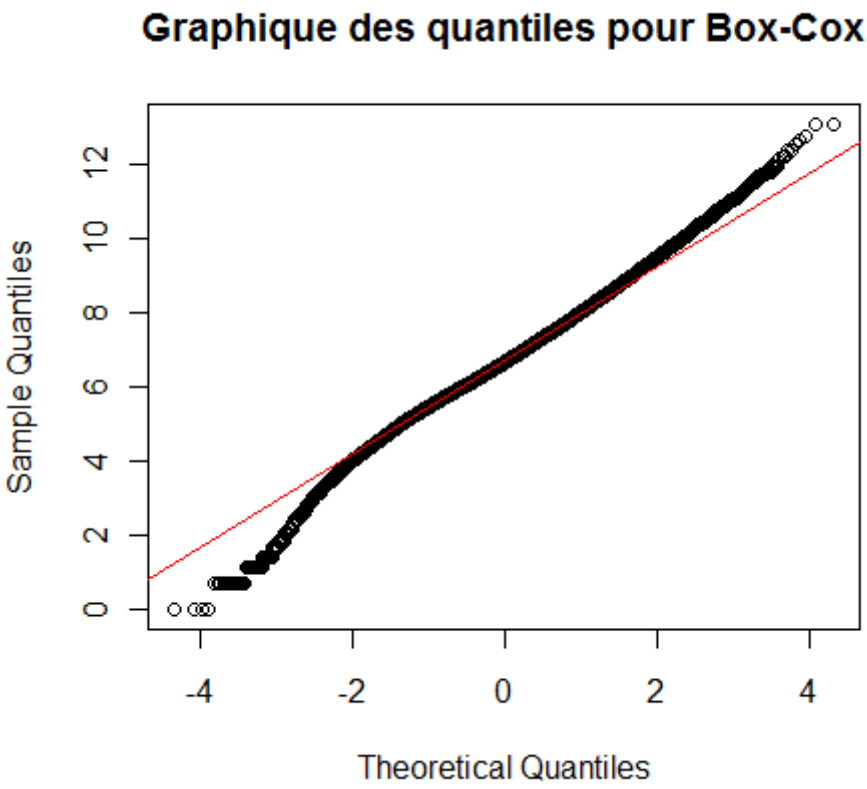
La transformation de Box-Cox pour ces données est donc :

$$y' = \frac{(y^{0,4} - 1)}{0,4}$$

D'après les représentations graphiques ci dessous nous observons une normalisation des données.







Nous allons utiliser un test de D'Agostino pour tester la normalité de nos nouvelles données.

```
> dagoTest(Boxdonnees)

Title:

 D'Agostino Normality Test

Test Results:

  STATISTIC:

    Z3  | Skewness: -0.7185

  P VALUE:

    Skewness Test: 0.4724
```

La p-valeur est égale à 0.4724 ce qui implique le non-rejet de l'hypothèse de normalité. Les données transformées par la méthode de Box-Cox suivent donc une loi normale.

Conclusion : A l'aide de la transformation de Box-Cox, j'ai obtenu des données qui suivent une loi normale. Il est cependant nécessaire de vérifier l'homoscédasticité de nos nouvelles données, car cette hypothèse est nécessaire dans l'utilisation des tests paramétriques de Student ou encore d'analyse de la variance (anova).

B. Test d'homoscédasticité

Pour tester l'homoscédasticité des données j'ai utilisé le test de Bartlett.

```
> bartlett.test(Boxdonnees,Orientation2)

        Bartlett test of homogeneity of variances

data:  Boxdonnees and Orientation2

Bartlett's K-squared = 1044.892, df = 3, p-value < 2.2e-16

> bartlett.test(Boxdonnees,InTGA2)

        Bartlett test of homogeneity of variances

data:  Boxdonnees and InTGA2

Bartlett's K-squared = 29.0555, df = 1, p-value = 7.034e-08

> bartlett.test(Boxdonnees,Species2)

        Bartlett test of homogeneity of variances

data:  Boxdonnees and Species2

Bartlett's K-squared = 1053.793, df = 10, p-value < 2.2e-16
```

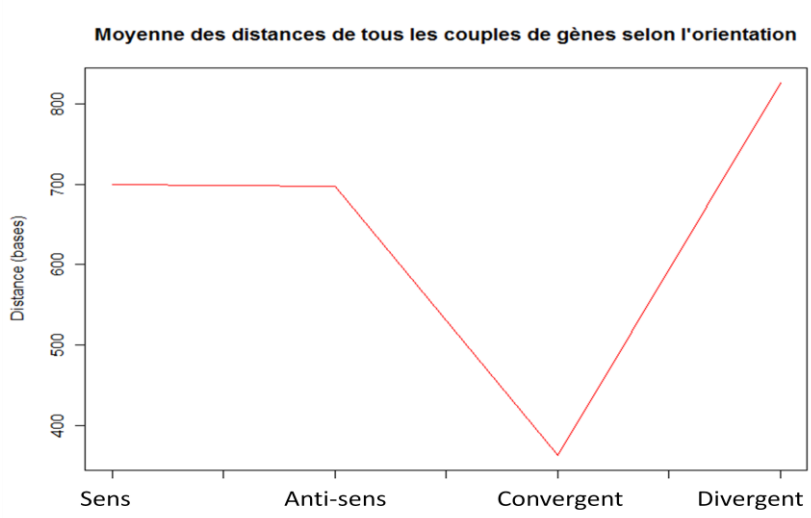
Les p-valeurs obtenues sont inférieures à 5%, ce qui traduit que l'hypothèse d'homoscédasticité est rejetée et donc qu'il n'y a pas d'égalité des variances. Il est donc nécessaire d'effectuer des tests non paramétriques au cours de notre étude avec les données de départ.

V. Etude statistique

Je vais tout d'abord rappeler la problématique. Il s'agit de déterminer si il existe une corrélation entre différentes caractéristiques d'un couple de gènes comme la distance intergénique, l'orientation des gènes, le fait d'être en tandem ou encore l'espèce.

A. Effet de l'orientation des gènes sur les distances intergéniques ?

Notre première approche a été d'étudier s'il existe un effet significatif de l'orientation sur les distances intergéniques tout gènes confondus (tandem et hors tandems). Il s'agit donc de comparer les moyennes des distances selon les quatre orientations : sens, anti-sens, convergente et divergente.



Nous observons que les moyennes des distances intergéniques en orientation sens et anti-sens semble être égales, que la moyenne des distances en orientation convergente semble être inférieurs aux deux précédentes et que la moyenne des distances en orientation divergente semble être supérieurs aux trois précédentes.

J'ai utilisé un test non paramétrique de Kruskal-Wallis pour comparer les quatre moyennes des distances selon leur orientation. Je vais commencer par présenter ce test avant d'énoncer les résultats obtenus.

Principe du test de Kruskal-Wallis

Le test de Kruskal-Wallis est un test non-paramétrique. Il s'applique sur des données quantitatives pour comparer les moyennes de plusieurs échantillons indépendants. L'unique hypothèse à vérifier pour utiliser ce test est l'indépendance des échantillons. Le test s'effectue en plusieurs étapes qui seront illustrées à l'aide d'un exemple :

Etape 1/ Classer les données sous forme de tableau.

Noter l'effectif de chaque série.

Exemple pratique : On veut comparer 3 milieux de culture différents A, B et C. Pour cela on compte le nombre de colonies bactériennes dans chaque milieu sur plusieurs jours.

Milieu	J1	J2	J3	J4	J5	J6
A	7	4	3	2	4	-
B	5	4	4	1	3	5
C	6	7	6	5	7	6

On obtient na=5, nb=6 et nc=

Etape 2/ Ranger les données en fonction de leur fréquence dans chaque série.

Dans notre série :

Nombre de colonies	1	2	3	4	5	6	7
Fréquence dans A	0	1	1	2	0	0	1
Fréquence dans B	1	0	1	2	2	0	0
Fréquence dans C	0	0	0	0	1	3	2

Etape 3/ Calculer la somme des fréquences.

Dans notre exemple :

Nombre de colonies	1	2	3	4	5	6	7
Fréquence dans A	0	1	1	2	0	0	1
Fréquence dans B	1	0	1	2	2	0	0
Fréquence dans C	0	0	0	0	1	3	2
Somme des fréquences	**1**	**1**	**2**	**4**	**3**	**3**	**3**

Etape 4/ Classer les données en rang par ordre.

Dans notre exemple :

Nombre de colonies	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Fréquence dans A	0	1	1	2	0	0	1
Fréquence dans B	1	0	1	2	2	0	0
Fréquence dans C	0	0	0	0	1	3	2
Somme des fréquences	**1**	**1**	**2**	**4**	**3**	**3**	**3**
Rang	**1**	**2**	**3** **4**	**5** **6** **7** **8**	**9** **10** **11**	**12** **13** **14**	**15** **16** **17**

Etape 5/ Calculer le rang corrigé qui est la moyenne des rangs pour chaque fréquence Rc

Nombre de colonies	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Fréquence dans A	0	1	1	2	0	0	1
Fréquence dans B	1	0	1	2	2	0	0
Fréquence dans C	0	0	0	0	1	3	2
Somme des fréquences	**1**	**1**	**2**	**4**	**3**	**3**	**3**
Rang	**1**	**2**	**3** **4**	**5** **6** **7** **8**	**9** **10** **11**	**12** **13** **14**	**15** **16** **17**
Rang corrigé	**1**	**2**	**3,5**	**6,5**	**10**	**13**	**16**

Etape 6/ Calculer les fréquences corrigées fc

fc = f x Rc

Nombre de colonies	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Fréquence dans A	0	1	1	2	0	0	1
Fréquence dans B	1	0	1	2	2	0	0
Fréquence dans C	0	0	0	0	1	3	2
Somme des fréquences	**1**	**1**	**2**	**4**	**3**	**3**	**3**
Rang	**1**	**2**	**3** **4**	**5** **6** **7** **8**	**9** **10** **11**	**12** **13** **14**	**15** **16** **17**
Rang corrigé	**1**	**2**	**3,5**	**6,5**	**10**	**13**	**16**
Fc pour A	**0**	**2**	**3,5**	**13**	**0**	**0**	**16**
Fc pour B	**1**	**0**	**3,5**	**13**	**20**	**0**	**0**
Fc pour C	**0**	**0**	**0**	**0**	**10**	**39**	**32**

Etape 7/ Calculer le total des rangs :

Ri = S Fc

Dans notre exemple :

Ra = 0+2+3.5+13+0+0+16 = 34,5

Rb = 1+0+3.5+13+20+0+0 = 37,5

Rc = 0+0+0+0+10+39+32 = 81

Etape 8/ Calcul de H :

$$H = \frac{12}{N \times (N+1)} \times \sum \frac{R_j^2}{n_j} - 3 \times (N+1)$$

N étant l'effectif total (somme de na, nb et nc)
R étant le total des rangs corrigés
n étant l'effectif de chaque série

Dans notre exemple :

na = 5 Ra = 34,5
nb = 6 Rb = 37,5
nc = 6 Rc = 81

$$H = \left(\frac{12}{17 \times 18}\right) \times \left[\frac{(34.5)^2}{5} + \frac{(37.5)^2}{6} + \frac{(81)^2}{6}\right] - 3 \times 18 = 7.40$$

Etape 9/ Calcul de la correction C

$$C = 1 - \frac{\sum T}{N^3 - N}$$

$$T = t^3 - t$$

t étant le nombre d'ex aequo

Dans notre exemple :

Dans la série A le nombre d'ex aequo t = 1
Dans la série B le nombre d'ex aequo t = 2
Dans la série C le nombre d'ex aequo t = 2
Donc : Ta=0 Tb=6 Tc=

$$C = 1 - \frac{6 + 6 + 0}{17^3 - 17} = 0.9975$$

Etape 10/ Calculer H'

$$H' = \frac{H}{C}$$

Dans notre exemple :

$$H' = \frac{7.40}{0.9975} = 7.418$$

Etape 11/Si tous les effectifs sont supérieurs ou égaux à 5, comparer H' avec la valeur du χ^2 pour (k-1) degré de liberté, k étant le nombre d'échantillons.

Si H' est supérieur au χ^2 de la table, il existe donc une différence significative entre les séries.
Si H' est inférieur au χ^2 de la table, il n'existe pas de différence significative entre les séries.

Dans notre exemple : Pour k-1 =2, degré de liberté la table du χ^2 montre 5,99 H' = 7.418
Donc H' est supérieur à la valeur du χ^2 lue, il existe ainsi une différence significative entre les 3 milieux de culture.

Etape 12/ Si un ou plusieurs effectif est inférieur à 5, on utilise la table de Kruskal Wallis qui donne les valeurs de H théorique.

Si H est inférieur au seuil de la table, il n'existe pas de différence significative entre les séries. Si H est supérieur au seuil de la table, il existe une différence significative entre les séries.

Application du test de Kruskal-Wallis sur nos données

J'ai commencé par vérifier l'indépendance de nos quatre groupes (4 orientations) qui est une hypothèse nécessaire dans l'utilisation du test de Kruskal-Wallis. L'indépendance des quatre groupes est justifiée par le contexte biologique. Il n'y a aucun lien entre les distances de deux groupes différents.

```
> kruskal.test(Distance2,Orientation2)

        Kruskal-Wallis rank sum test

data:  Distance2 and Orientation2

Kruskal-Wallis chi-squared = 11241.04, df = 3, p-value < 2.2e-16
```

Avec nos données le test de Kruskal-Wallis renvoie une p-valeur inférieure à 2.2*10^-16. On rejette donc l'hypothèse de départ, c'est-à-dire, l'égalité des quatre moyennes et l'on conclut que toutes les moyennes ne sont pas égales. Il est donc intéressant d'identifier quelles différences entre les moyennes sont significatives. Pour cela, j'ai testé les moyennes deux à deux en utilisant le test non-paramétrique de **Wilcoxon-Mann-Whitney**. Ce test est un cas particulier du test de Kruskal-Wallis. Il s'agit du test de Kruskal-Wallis dans le cas de deux échantillons. Ces hypothèses sont les mêmes que pour le test de Kruskal-Wallis, c'est-à-dire que les échantillons doivent être de nature quantitative et indépendants.

J'ai donc testé l'égalité de chacune de nos moyennes avec les moyennes des autres groupes. Ayant quatre groupes, cela revient à utiliser six tests de Wilcoxon-Mann-Whitey.

Exemples :

```
> wilcox.test(sens,convergent)

        Wilcoxon rank sum test with continuity correction

data:  sens and convergent

W = 201517609, p-value < 2.2e-16

> wilcox.test(sens,antisens)

        Wilcoxon rank sum test with continuity correction

data:  sens and antisens

W = 113780139, p-value = 0.9482
```

J'ai obtenu des p-valeurs inférieurs à 0.05 dans le cas des tests entre les moyennes des distances d'orientations en sens contre convergent, sens contre divergent, anti-sens contre convergent, anti-sens contre divergent et convergent contre divergent. Cela implique que les différences de moyenne sont significatives dans les cinq cas précédents. On remarque que la différence entre sens et anti-sens n'est pas significative.

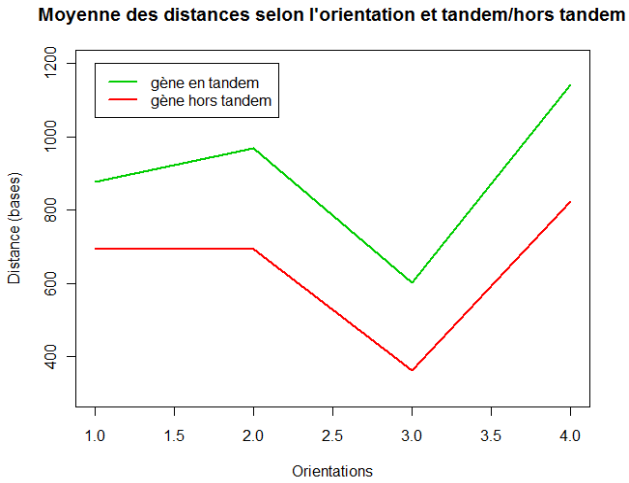
Conclusion :

Les tests statistiques réalisés confirme donc bien les observations faites préalablement : la moyenne des distances en orientation convergente est inférieure aux moyennes en sens et en anti-sens qui elles-mêmes sont inférieures à la moyenne des distances en orientation divergente. Cependant il n'y a pas de différence significative des moyennes des distances en orientation sens et en orientation anti-sens.

B. Effet du fait d'être en tandem sur les distances intergéniques ?

La deuxième question que l'on se pose est la suivante : existe-t-il une différence significative des distances en fonction de leur orientation entre les tandems (InTGA=1) et les hors tandem (InTGA=0) ?

Ici les orientations en sens, anti-sens, convergent et divergent sont respectivement notés par 1, 2, 3 et 4.



Nous observons sur le précédent graphique que le fait d'être en tandem pour un couple de gène ne semble pas changer le profil des moyennes des distances selon l'orientation.

Dans un premier temps, on compare les moyennes des deux échantillons (InTGA=1 et InTGA=0). Nos données comportent 591 valeurs de distances pour les gènes en tandem et 65171 valeurs pour les gènes hors tandem. Nos données ne satisfont pas à l'hypothèse d'homoscédasticité. Donc comme pour la question précédente, on se résout à utiliser des tests non-paramétriques. Le contexte nous permet à nouveau de conclure quant à l'indépendance entre les distances des gènes en tandem et celles des gènes non en tandem.

Chacun de nos échantillons possède quatre groupes : un groupe par orientation. J'ai donc utilisé un test de Kruskal-Wallis pour comparer les quatre moyennes de distance de chaque échantillon. La p-valeur est inférieure à 0.05 dans le cas des tandems et des hors tandem. Nous en déduisons une différence significative entre les quatre moyennes pour chacun de ces échantillons. Pour avoir plus de détails on effectue des tests de Wilcoxon-Mann-Whitney deux à deux sur les groupes de chaque échantillon. Les tests sont effectués comme précédemment pour identifier des différences significatives de moyennes de distances en comparant chaque groupe avec les autres au sein de chaque échantillon. Par exemple : nous comparons à l'aide d'un test de Wilcoxon-Mann-Whitney le groupe constitué des distances des gènes en tandem dans l'orientation sens avec le groupe des distances des gènes en tandem avec l'orientation anti-sens.

Les résultats observés dans le cas des distances des hors tandem est le même que dans le cas où toutes les distances avaient été utilisés, ce qui traduit le fait que les hors tandems représentent la majorité des données étudiées. Ainsi, nous remarquons des différences significatives des moyennes des distances selon l'orientation lors des tests de l'orientation sens contre convergent, sens contre divergent, anti-sens contre convergent, anti-sens contre divergent et convergent contre divergent.

Dans le cas des tandems, nous ne retrouvons pas les mêmes résultats. Les différences des moyennes des distances selon l'orientation sont significatives lors des tests de sens contre anti-sens, sens contre divergent, anti-sens contre convergent, anti-sens contre divergent et convergent contre divergent. Il n'y a pas de différence significative des moyennes des distances selon l'orientation en sens contre convergent. Ceci s'explique par le faible nombre d'éléments dans deux des groupes (convergent et divergent). Pour les gènes en tandems voici un tableau qui résume l'effectif de chacun des groupes d'orientation et la moyenne des distances :

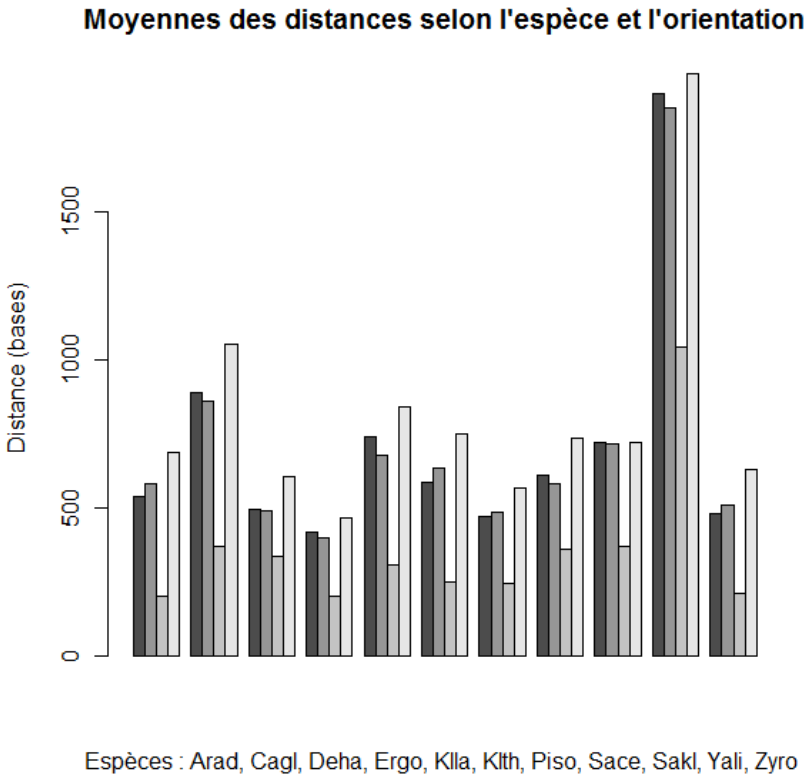
	Sens	Anti-sens	Convergent	Divergent
Effectif	232	251	47	61
Moyenne (bases)	876	968	601	1143

Le tableau explique certains des résultats obtenus à l'aide des tests de Wilcoxon-Mann-Whitney. En effet, la différence des distances moyennes entre sens et convergent (275 bases) est plus grande qu'entre sens et anti-sens (92 bases) mais cette différence n'est pas significative tandis qu'elle l'est entre les deux orientations sens et anti-sens. Le nombre d'effectif dans chacun des groupes est très différent c'est pourquoi on arrive à des résultats aussi surprenant par rapport aux différences que l'on observe sur la totalité des données.

C. Effet de l'espèce sur les distances intergéniques dans chaque orientation

La troisième question posée était de savoir si les moyennes des distances en fonction de l'orientation dans chacune des espèces étaient significativement différentes et par la suite de savoir si elles avaient le même comportement que dans l'ensemble des espèces. Le but étant d'identifier si une espèce pouvait « fausser » les résultats, ou possédait des caractéristiques spécifiques.

Dans un premier temps il s'agissait de traiter cette question de manière encore plus précise : est-ce que les moyennes des distances en fonction de l'orientation **et du fait d'être en tandem** dans chacune des espèces étaient significativement différentes et par la suite de savoir si elles avaient le même comportement que dans l'ensemble des espèces. Cependant au vu du nombre très faibles d'éléments présents dans nos groupe de l'échantillon des tandems, il a été décidé de supprimer le fait d'être en tandem dans la question afin d'augmenter la taille de nos groupes.



Orientation :

: Sens : Anti-sens :Convergent : Divergent

Toutes les espèces semblent avoir le même profil général, excepté Sakl ou les moyennes des distances selon l'orientation sens, anti-sens et divergent ne semblent pas être significativement différentes. De plus, l'espèce Yali a des distances moyennes supérieures aux autres espèces.

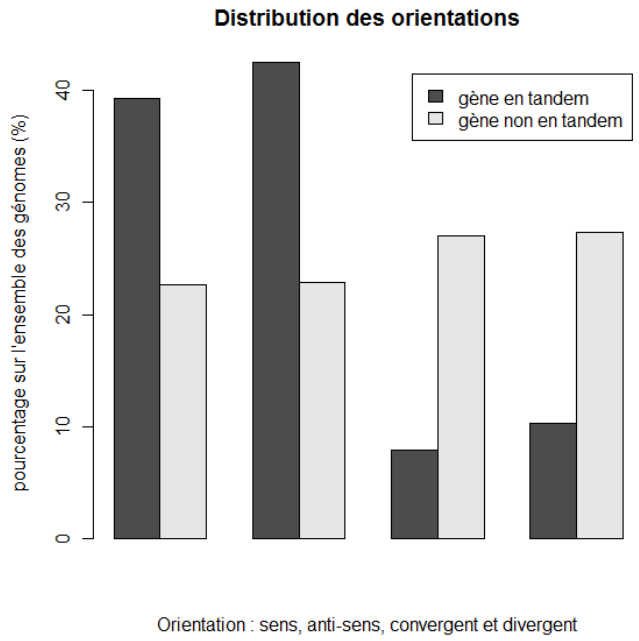
Nous nous intéresserons pour le moment qu'au profil général des espèces. Il s'agit d'utiliser des tests d'égalité des moyennes dans chacune de nos espèces. J'ai utilisé le test de Kruskal-Wallis pour comparer les moyennes des distances selon l'orientation dans chacune des espèces. Par la suite, on utilisera des tests de Wilcoxon-Mann-Whitney pour déterminer quelles différences sont significatives deux-à-deux dans les espèces où le test de Kruskal-Wallis nous a rejeté l'hypothèse d'égalité des moyennes.

J'ai effectué les tests de Kruskal-Wallis dans chacune des espèces et il s'est révélé que toutes les espèces obtiennent le même résultat. Chaque espèce rejette l'hypothèse d'égalité des moyennes selon l'orientation. J'ai donc effectué six tests de Wilcoxon-Mann-Whitney dans chacune des espèces.

Après avoir réalisé l'ensemble des tests on observe trois types d'espèce qui réagissent de manières différentes. Les espèces Arad (#1), Deha (#3), Ergo (#4), Klla (#5), Klth (#6), Piso (#7), Sace (#8), Yali (#10) et Zyro (#11) ont toutes le même profil général : pas de différence significative entre les orientations sens et anti-sens, mais des différences significatives dans tous les autres cas. L'espèce Cagl (#2) a un profil légèrement différent : la différence de moyenne de distance n'est pas significative entre les orientations sens contre anti-sens et sens contre divergent. Pour finir l'espèce Sakl (#9) possède un profil qui lui est propre. Les différences de moyennes ne sont pas significatives entre les orientations sens contre anti-sens et anti-sens contre divergent.

Conclusion : il existe trois profils différents d'espèces quand on regarde la significativité des différences entre les moyennes des distances selon l'orientation. On remarque cependant que dans chaque espèce, la différence entre l'orientation sens contre anti-sens n'est jamais significative et que les différences entre les orientations sens contre convergent, anti-sens contre convergent et convergent contre divergent sont quant à elles toujours significatives. Donc les moyennes des distances en orientation convergente sont toujours significativement différentes des moyennes dans une autre orientation.

D. Existe-t-il une différence significative concernant les distributions des gènes en tandem et hors-tandem en fonction de l'orientation ?



Dans le cas des gènes en tandem, il semble que les orientations directes (sens et anti-sens) sont privilégiées. De plus il ne semble pas y avoir de différence significative du pourcentage entre les orientations directes ou indirectes et ce également dans le cas des gènes hors tandem. Dans le cas des gènes hors tandem, les orientations indirectes semblent être très légèrement privilégiées.

J'ai réalisé une table de contingence de l'orientation et du fait d'être en tandem afin étudier les profils des orientations en fonction du fait d'être en tandem (0 = hors-tandem, 1= tandem).

```
> table<-table(Orientation2,InTGA2)
> table
```

	InTGA2	
Orientation2	0	1
1	14791	232
2	14903	251
3	17649	47
4	17828	61

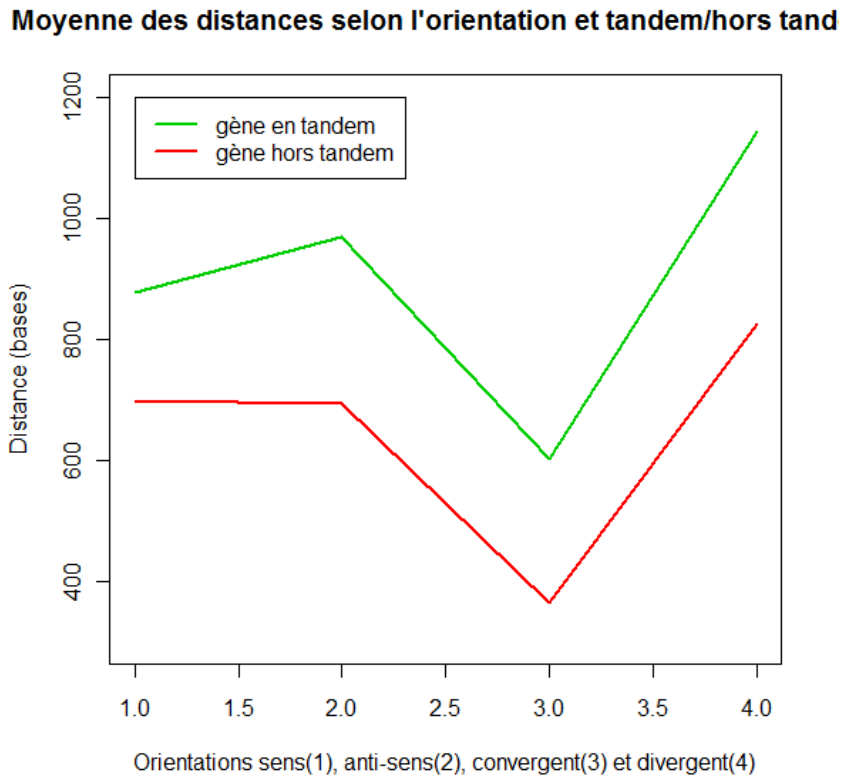
J’ai ensuite effectué un test du Chi-deux pour mettre en évidence l’éventuelle indépendance des deux variables :

```
> chisq.test(table)

        Pearson's Chi-squared test

data:  table

X-squared = 310.0882, df = 3, p-value < 2.2e-16
```

Avec une p-valeur inférieure à 0.05, nous en déduisons que les distributions des tandems/hors tandems sont significativement différentes en fonction de l’orientation.

En testant deux à deux les distributions selon l’orientation, toujours à l’aide du test du Chi-deux, on remarque que l’orientation sens et anti-sens ont le même profil et également que les orientations convergentes et divergentes ont le même profil, qui est différent du précédent.

E. Comparaison des moyennes des distances des gènes en tandem et des gènes hors-tandem en fonction de l’orientation



Ici, nous nous intéressons à l’écart entre les moyennes des distances pour les gènes en tandem et les gènes hors-tandem. On s’intéresse tout d’abord à la différence des distances uniquement en fonction du fait d’être en tandem (InTGA). On réalise un test de comparaison des moyennes de Wilcoxon-Mann-Whithney. D’après ce test, on retrouve une p-valeur inférieure à 2.2*10^-16. Nous rejetons donc l’hypothèse nulle : les moyennes des distances des tandems et des hors-tandems sont égales. Nous acceptons donc l’hypothèse alternative : les moyennes des distances des tandems et des hors tandem sont différentes.

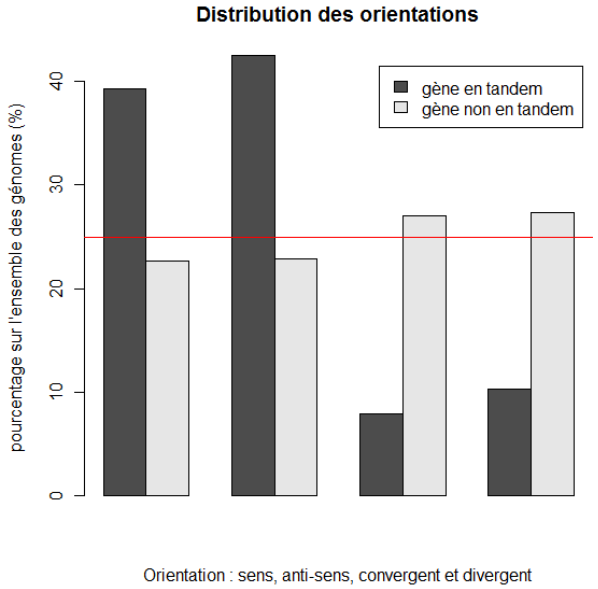
Il est ensuite intéressant d'étudier si la différence observée est due à une orientation particulière ou si l'on retrouve ces différences sur les quatre orientations. Pour cela on effectue des tests non paramétriques de Wilcoxon-Mann-Whitney, pour comparer les moyennes des distances des tandems et des hors-tandem selon leur orientation. Les résultats pour les tests de Wilcoxon-Mann-Whitney nous indiquent que les différences sont significatives pour toutes les orientations.

Conclusion :

L'écart observé sur le graphique entre la distance moyenne selon les orientations entre les gènes en tandem et les gènes hors-tandem est donc significatif.

F. Proportion de tandems/ hors tandems selon l'orientation

Finalement, la répartition des gènes en tandem et des gènes non en tandem est-elle significativement différente d'une répartition uniforme, c'est-à-dire 25% de gènes dans chaque orientation. Cette question est intéressante car il a été montré que la répartition des gènes sur le brin sens et le brin anti-sens est non privilégié (50% sur chaque brin). On pourrait donc s'attendre à des proportions de 25% pour chacune des orientations.



Le trait rouge représente le niveau des 25% attendu dans chacune des orientations

J'ai comparé les effectifs observés dans nos 8 groupes (selon les quatre orientations et le fait d'être en tandem qui a deux modalités) avec les effectifs attendus dans nos groupes. J'ai utilisé un test du Chi-deux pour comparer dans le cas de gènes en tandem et de cas de gènes non en tandem.

Les résultats sont les suivants :

Dans le cas des gènes en tandem :

```
> mat1<-matrix(c(232,251,47,61,591/4,591/4,591/4,591/4), nrow=4)
> mat1
     [,1]   [,2]
[1,]  232 147.75
[2,]  251 147.75
[3,]   47 147.75
[4,]   61 147.75
> chisq.test(mat1)

        Pearson's Chi-squared test

data:  mat1
X-squared = 133.5979, df = 3, p-value < 2.2e-16
```

Dans le cas des gènes non en tandem :

```
> mat2<-matrix(c(14791,14903,17649,17828,65171/4,65171/4,65171/4,65171/4),nrow=4)
     [,1]     [,2]
[1,] 14791 16292.75
[2,] 14903 16292.75
[3,] 17649 16292.75
[4,] 17828 16292.75
> chisq.test(mat2)

        Pearson's Chi-squared test

data:  mat2
X-squared = 257.7378, df = 3, p-value < 2.2e-16
```

Conclusion : on en déduit d'après les tests du Chi-Deux que les répartitions observées sont significativement différentes des répartitions uniformes.

Dans le cas des tandems : les proportions des orientations sens et anti-sens sont supérieures à la moyenne théorique tandis que les proportions des orientations convergente et divergente sont inférieures à la moyenne théorique.

Dans le cas des hors-tandems : on observe le contraire, les proportions des orientations sens et anti-sens sont inférieures à la moyenne théorique et les proportions des orientations convergente et divergente sont supérieures à la moyenne théorique.

VI. Conclusion

De nombreuses observations se sont révélées être significatives à l'issue des tests statistiques. Nous avons obtenu les résultats suivants :

- Les moyennes des distances intergéniques selon l'orientation des gènes sont significativement différentes dans l'ensemble des couples de gène pour les orientations en sens contre convergent, sens contre divergent, anti-sens contre convergent, anti-sens contre divergent et convergent contre divergent. Il n'y a pas de différence significative des moyennes des distances pour l'orientation en sens contre anti-sens.
- Dans le cas des hors-tandem nous observons le même résultat que pour l'ensemble des gènes
- Dans le cas des gènes en tandem nous observons des différences significatives des moyennes des distances selon l'orientation sens contre anti-sens, sens contre divergent, anti-sens contre convergent, anti-sens contre divergent et convergent contre divergent. Il n'y a pas de différence significatives entre les moyennes des distances en orientation sens contre convergent
- Nous observons trois types de profil d'espèce différents en identifiant les différences significatives des moyennes de distances en fonction de l'orientation au sein de chaque espèce. L'espèce Cagl et Sakl possèdent chacune un profil qui lui est propre tandis que les autres espèces ont toutes le profil général (celui obtenu lors de l'étude sur toutes les espèces)
- La distribution des orientations dans le cas des tandems favorise les orientations directes (sens et anti-sens) qui possèdent près de 40% de présence chacune tandis que les orientations inverses (convergent et divergent) se partage équitablement les 20% restant.
- Il existe une différence significative des moyennes des distances intergéniques du fait d'être en tandem ou en hors tandem. Les distances pour les gènes en tandem sont supérieures à celles des gènes hors tandem quelquesoit l'orientation

En guise de conclusion personnelle sur ce stage de fin de M1, j'ajouterais que ce fut très instructif et un plaisir de travailler au sein de l'équipe de l'UMR 7156. Ce stage de deux mois m'a permis de confirmer mon projet professionnel dans le domaine de la statistique