

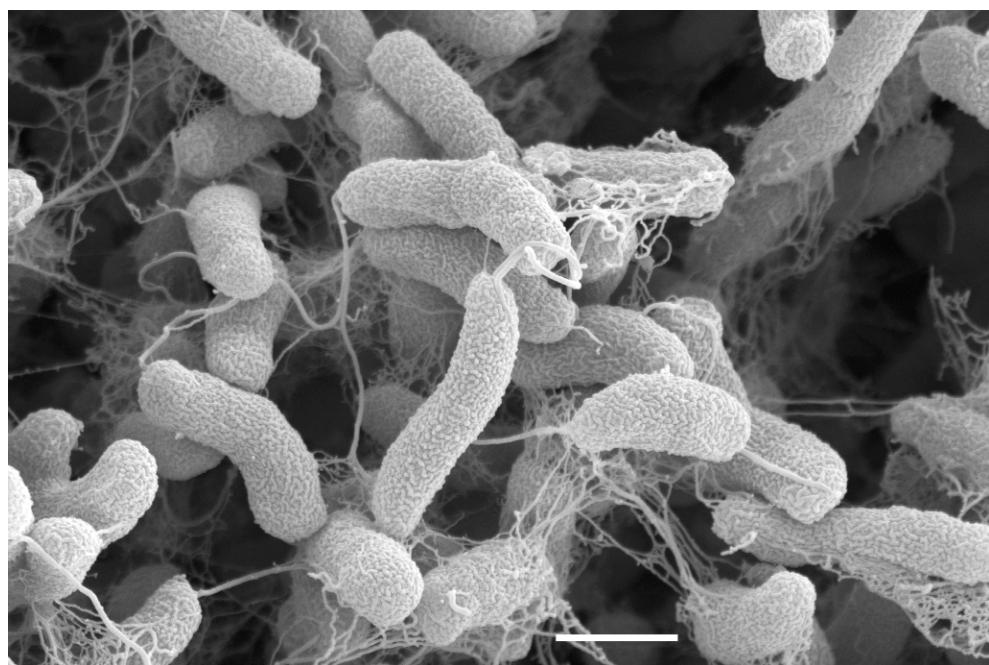
Projet : Analyse de génome

Analyse de manière comparative de *E.coli* et de *V.cholerae*

Lachiheb Sarah et Leroy Cassandre - 6 avril 2017



E.coli

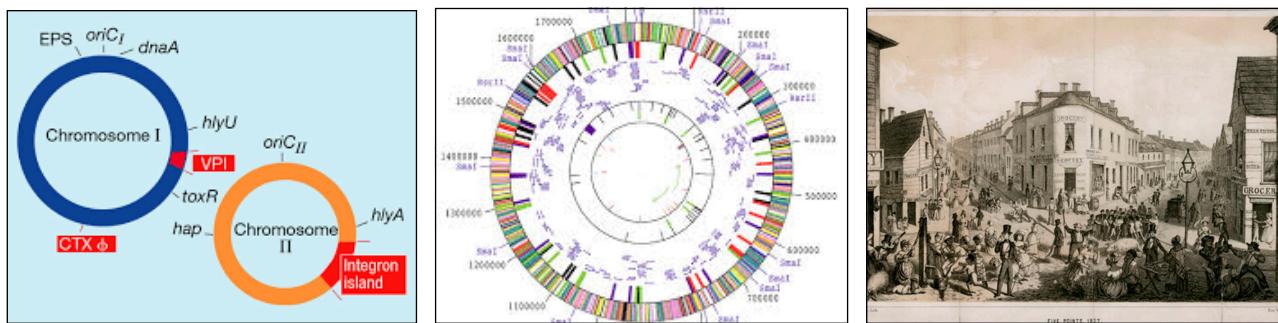


V.cholerae

Introduction

Le choléra est une maladie infectieuse très contagieuse. Elle est due à *Vibrio Cholerae* (*V.Cholerae*) une bactérie à bacilles Gram négatif. Plus de 200 groupes ont été défini, dont seulement deux, O1 et 0139 sont susceptibles de sécréter la toxine cholérique et donc de provoquer le choléra. Ici nous étudions O1.

A quoi est due la pathogénicité virulente de ces deux bactéries uniquement ?



Chromosomes de *V.Cholerae*, Chromosome de *E.Coli*, Epidémie de Choléra en 1832 à New York

Les procaryotes sont des cellules hyper-diversifiées capables de coloniser les territoires les plus hostiles, « Si elles ont longtemps été considérées comme les briques élémentaires du vivant, ont les voit aujourd’hui comme des organismes génétiquement composite, champion du vol de gène ». Un gène est une séquence d’ADN (Acides DésoxyriboNucléiques) qui spécifie la synthèse spécifiant un caractère (protéines, acide ribonucléique...)

Elles possèdent donc une grande capacité à intégrer du matériel génétique d’autre espèces, peu ou très éloignées.

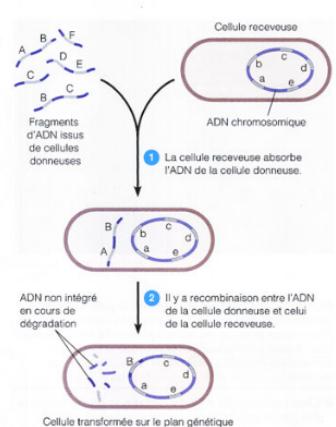
Ce mécanisme se nomme le transfert horizontal de gènes par opposition au transfert vertical qui est la transmission de la mère à la fille.

Le cytoplasme de la bactérie, où se trouve le génome, est efficacement protégé du reste par une membrane. La traversée, pour l’ADN issue de l’extérieur est le résultat de mécanismes spéciaux.

Une fois l’ADN étranger entré, il peut y avoir plusieurs possibilités:

- Il peut être détruit par le système de dégradation de l’ADN (enzyme ...)
- Ou se maintenir en tant qu’entité réplicative autonome comme les plasmides (plasmide que ne possède pas *V.Cholerae*).
- Sinon il peut, pour tout ou partie, être intégré au génome de l’hôte. Dans ce cas, plusieurs sous cas se présentent, soit il est intégré en tant que partie d’ADN similaire à l’hôte et donc, pourra remplacer des sections homologues préexistantes par des nouvelles. Ou alors,

Transfert Horizontal:



s'il sont trop différents, simplement ajouter de nouveaux gènes avec de nouvelles fonctionnalités.

Cela fait indéniablement partie du processus d'évolution, et n'a été exploité que très récemment par le séquençage de génome. Un génome est l'ensemble du matériel génétique d'une espèce codé dans son ADN. Ce dernier nous permet de démontrer qu'il peut y avoir des mouvements génétiques entre des espèces très éloignées, ce qui apparaît donc comme un moyen d'adaptation. On observe, par exemple, la capacité que montre certaines souches à acquérir des gènes de virulence, ou encore une résistance à un antibiotique ...

Afin d'identifier ces transferts horizontaux et donc de les détecter nous utiliserons dans un premier temps le pourcentage en G+C.

Le taux de GC, ou coefficient de Chargaff, d'une séquence d'ADN est défini comme la proportion de bases de cette séquence étant, soit une cytosine « C », soit une guanine « G ». L'adénine « A », et la thymine « T », ne rentrent pas dans ce calcul. Comme la guanine forme une liaison avec la cytosine, ce pourcentage permet aussi de calculer le pourcentage de liaison G-C dans l'ADN. En effet le transfert horizontal d'une partie du génome de bactérie, aux caractéristiques différentes, est donc identifiable par le taux en GC, qui sera potentiellement différent.

Aussi, il existe donc des îlots génomiques, souvent îlots de pathogénicité qui sont liés au pouvoir pathogène. Ces derniers correspondent à un ensemble de gènes présentant des caractères de virulence présent au niveau du chromosome, en un endroit. Ces séquences ont un pourcentage en G+C qui peut être différents du reste du génome et ce qui permet donc de le rendre facilement identifiable puisqu'il est acquis par transfert horizontal. Ceci pourraient expliquer que la majorité des souches de *Vibrio* ne soient pas pathogène, un transfert n'ayant peu être pas eu lieu les rends inoffensives.

Le but du projet est d'analyser les propriétés générales du génome, de détecter les différentes régions de composition homogène, de pouvoir trouver les éléments génétiques qui lui confèrent sa toxicité, et de comprendre comment cette toxicité a été acquise.

Ainsi dans un premier temps nous analyserons les propriétés globales de *V.Cholerae* et chercherons à détecter des hétérogénéités dans le génome séquencé via les fonctions faites lors des TME, des logiciels BLAST, COG et par visualisation.

Puis nous analyserons ces îlots de pathogénicité via le logiciel IslandViewer 3, ce qui nous permettra d'analyser les gènes pathogènes de *Vibrio Cholerae*.

Sommaire

Partie A : Propriété globales des génomes et détection des hétérogénéités dans la séquence.

I - Préliminaire et propriétés de bases.

A - Premières propriétés.

B - Les ORF's de Glimmer.

C - Les gènes.

II - Annotations par homologie avec BLAST.

A - Analyse des gènes codants.

B - A propos des non codants.

III - Analyse comparative, des propriétés et annotations des gènes correspondants

A- Assignation des catégories fonctionnelles pour les gènes codants avec COG.

B- Visualisation et navigation dans un génome avec analyse de composition.

Partie B : Annotation plus automatique des îlots de pathogénicité.

I - Les gènes pathogènes via IslandViewer 3.

II - Les compositions en codons.

Conclusion.

Partie A : Propriété globales des génomes et détection des hétérogénéités dans la séquence.

I - Préliminaire et propriétés de bases.

A- Premières propriétés.

	<i>E.coli</i>	<i>V.cholerae</i>
Nombre de chromosome	1, circulaire	2, circulaires
Nombre de plasmide	Aucun	Aucun
Longueur du ou des chromosomes	4 641 652 bp	2 961 152 bp pour le I 1 072 316 bp pour le II
Longueur du génome	Identique à la longueur de l'unique chromosome 4 641 652 bp	La somme des deux soit : 4 033 464 bp
Pourcentage en GC	51 %	47.49% Soit 47.7% pour le chromosome I et 46.9% pour le II
Compositions en nucléotides	A : 24.62% C : 25.42% G : 25.37% T : 24.59%	A : 26.11% C : 23.62% G : 23.86% T : 26.40%

Il est très utile de faire des calculs de pourcentage en nucléotides dans ce cas car le taux en GC est très utilisé en bactériologie, étude des bactéries, pour la **taxonomie**. Le but étant de regrouper les espèces du vivant en entités appelées taxons (familles, genres, espèces, etc...) afin de pouvoir les nommer et les classer. Cela est une méthode pour établir un **arbre phylogénétique**, comme ci dessous pour exemple, afin de comprendre d'où viens notre bactérie.

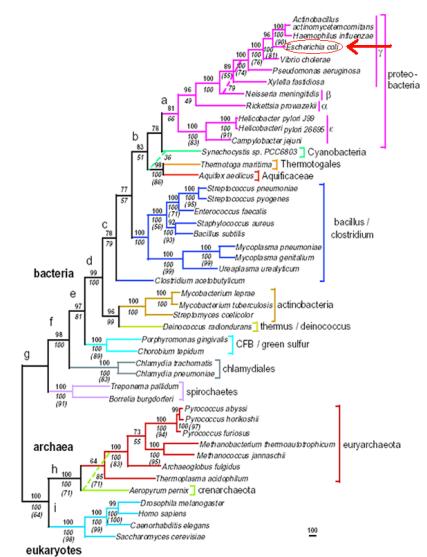
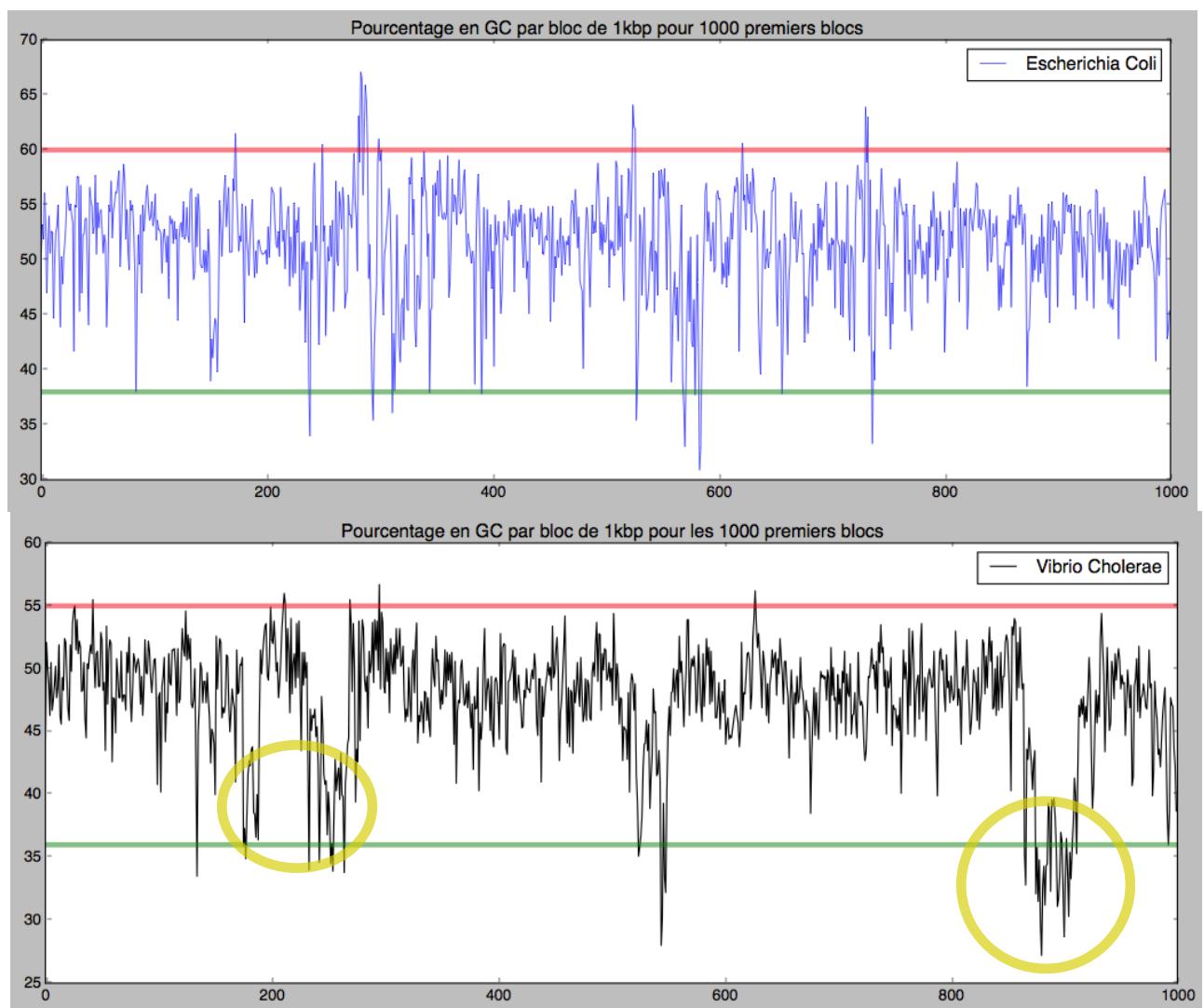


Figure 1:

Pour faciliter la compréhension de l'analyse du pourcentage de GC dans chaque groupe de 1000 nucléotides (1kbp) du génome, nous avons fait des courbes représentant leur pourcentage en G+C.

Distribution du pourcentage en GC dans des segments génomique de 1 kbp pour les nucléotides par tranche de 1 000 000 de nucléotides (1000 blocs de 1kbp) :



— Correspond à la limite où nous considérons que au-delà cela est un pic important du taux de GC.

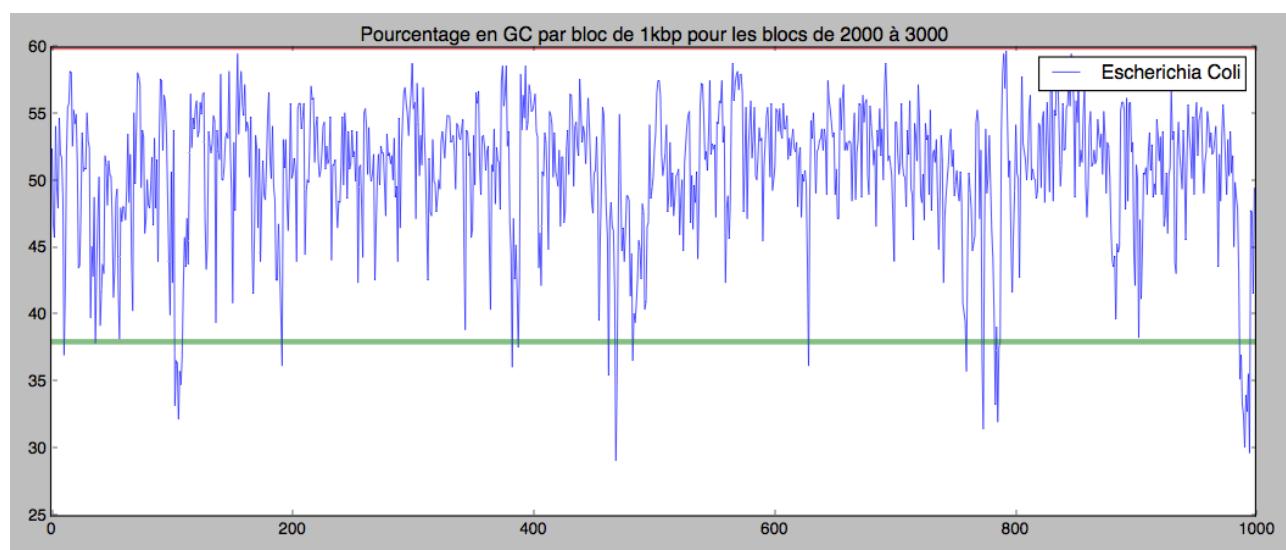
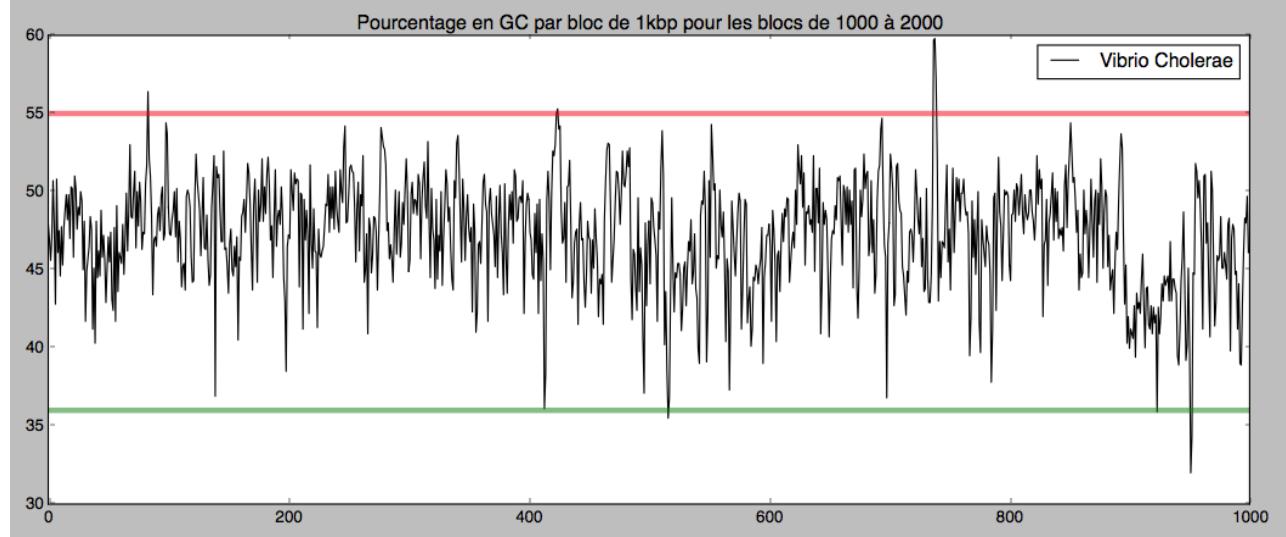
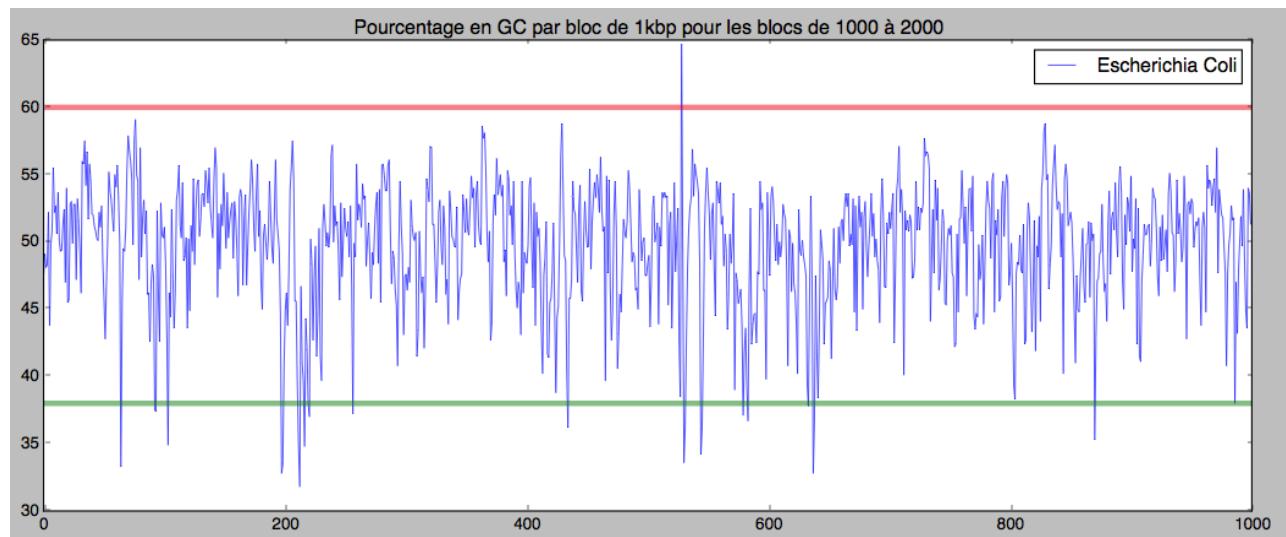
— Correspond à la limite où nous considérons que au-dessous cela est considéré comme une chute importante du taux de GC.

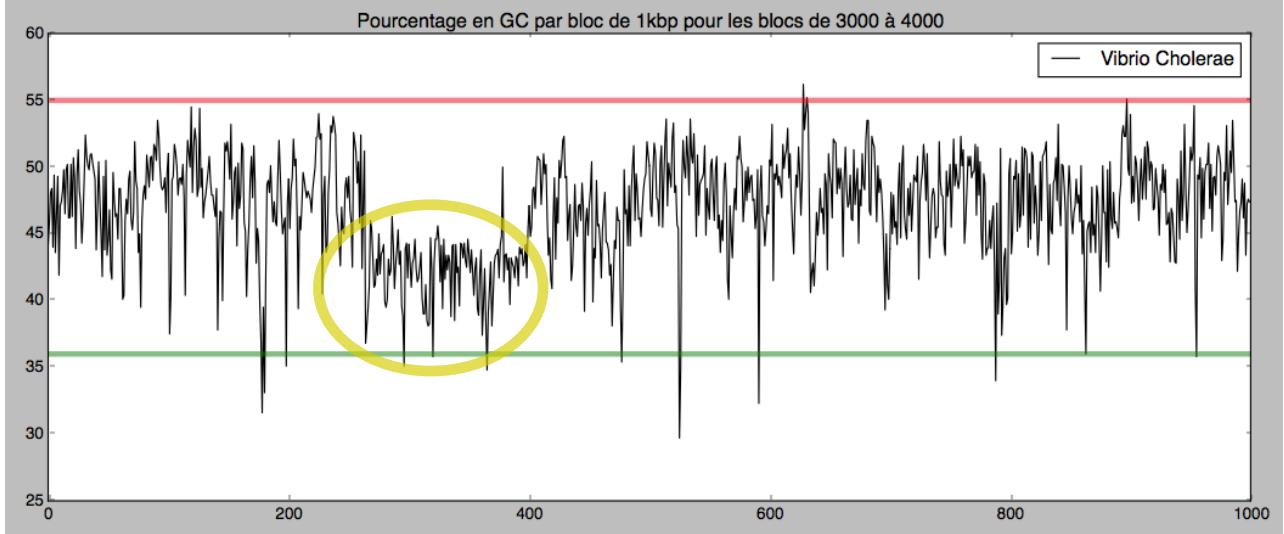
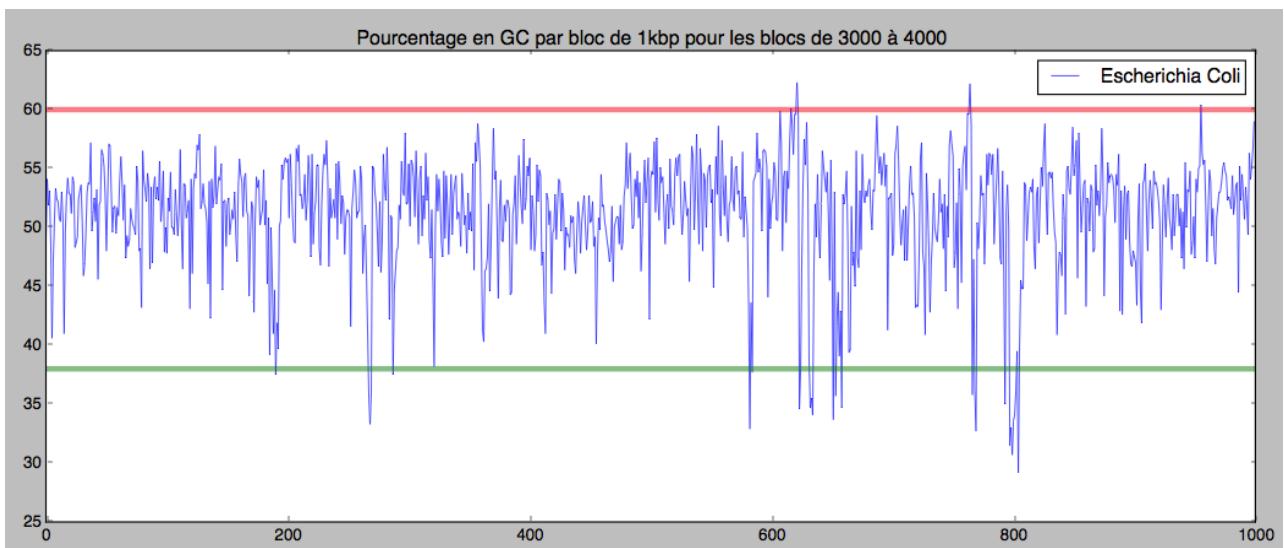
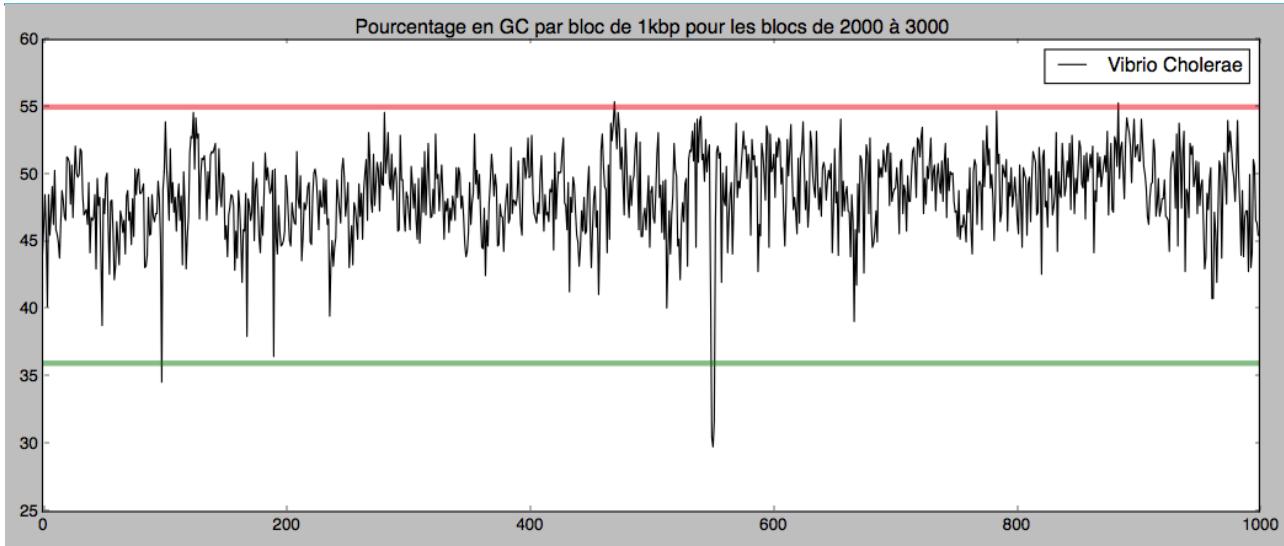


Anormalité prolongée.

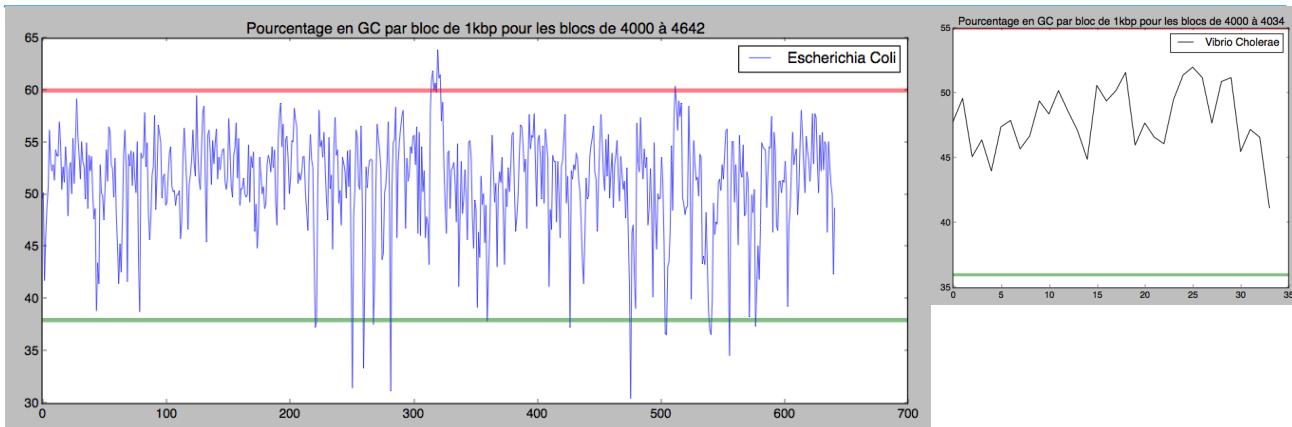
Etant donné que le pourcentage en GC de *V.Cholerae* est plus bas que pour *E.coli*, 47.5 contre 51, nous n'avons donc pas placé les limites rouges et verte aux même endroit pour

les deux génomes. La limite rouge est à 60 pour *E.Coli* contre 55 pour *Cholerae* et la verte à 38 pour *Coli* et 36 et *Cholerae*.





On remarque que le taux en GC est assez homogène malgré le fait que l'on peut observer, à certains endroits des pics extrêmes mais toujours en petit nombre, entre 1 et 3 Kbp. Cependant dans les 1000 premiers blocs du chromosome I on remarque des anomalies co-localisées (cercles jaunes) ainsi que dans le chromosome II ci dessus.



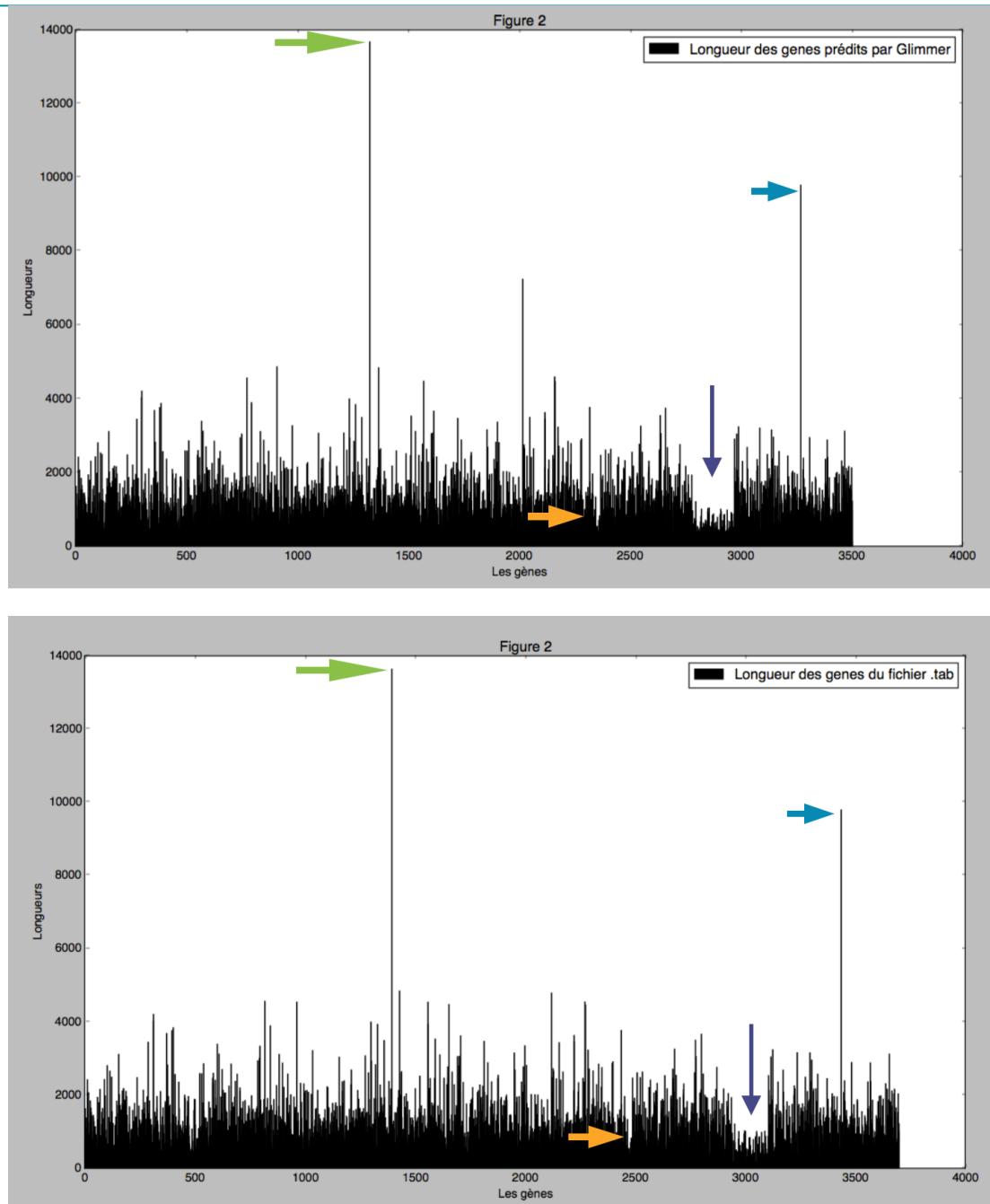
Nous savons que chaque génome à une **composition en G+C** qui lui est plus ou moins **propre** et qui reste **assez homogène**. Cependant nous trouvons une **régression**, en particulier, où la composition en nucléotides **contraste** avec la moyenne du génome, 35% contre 47.5% pour le reste. Ceci suggère donc d'après la définition donné en introduction que cette dernière à été acquise par transfert horizontal d'un génome ayant une compositions différente et donc un fort taux A-T. Souvent , les bactéries ayant un ancêtre commun proche ont un taux en GC similaire, donc nous considérerons que ce **transfert horizontal** à été fait avec une **espèce éloignée**.

On peut voir tout de suite que les anomalies sont plus présents dans la première moitié du génome. Elles ont tendance à être co-localisé, c'est à dire proches le uns des autres.

B- Les ORF's de Glimmer.

Nous avons fait deux histogrammes, le premier donnant les longueurs des gènes prédits par le site Glimmer, et le second ceux dans le fichiers .tab fournis.

Nous remarquons de grandes similitudes entre les deux histogrammes, notamment celles marquées par des flèches, qui sont facilement identifiables. Mais dès que l'on regarde plus attentivement nous remarquons qu'ils sembles réellement très proches en de nombreux points. Cependant, ils y a quelques différences, les ORF's de Glimmer sont plus nombreuses, 3650 contre 3504 pour les gènes présents dans le fichier .tab. Aussi la moyenne des tailles des gènes prédits par Glimmer est proche de la taille moyenne des gènes réelles, 938 nucléotides contre 979 en réalité, soit 41 nucléotides de moins pour Glimmer.



La matrice de confusion alors obtenue est, pour chaque chromosome :

Valeurs prédictives/ ORFs	Non gène	Gene
Non gène	429 754 (I) 185 349 (II) Vrai négatif	22 558 (I) 20 318 (II) Faux positif
Gène	1 312 219 (I) 414 971 (II) Faux négatif	1 233 633 (I) 465 081 (II) Vrai positif

La sensibilité calculée est alors de 0.48457 soit **48%** de probabilité qu'une ORF soit bien un gène pour le chromosome I et 0.52847 soit **53%** pour le second.

La spécificité de 0.95013, soit 95%, de probabilité qu'un non gène ne soit pas une ORF pour le premier et 0.90121, donc 90% pour le suivant.

et la valeur prédictive est donc de 0.98204 soit 98% de probabilité qu'un gène soit présent s'il détecte une ORFs le premier chromosome et une peu moins soit 96% pour le second.

On sait où sont les gènes et on mesure alors la capacité de Glimmer à prédire si est gène est présent ou non. Le résultat est donc bon car il est supérieur à **96% de chance d'obtenir un gène pour une ORF sélectionnée**. On peut en déduire cela car la spécificité et la sensibilité n'ont pas des valeurs trop éloignées, si la somme des deux faisait 100%, on ne pourrait rien déduire de nos résultats.

Ceci nous permet d'affirmer que Glimmer est un très bon prédicteur de gène.

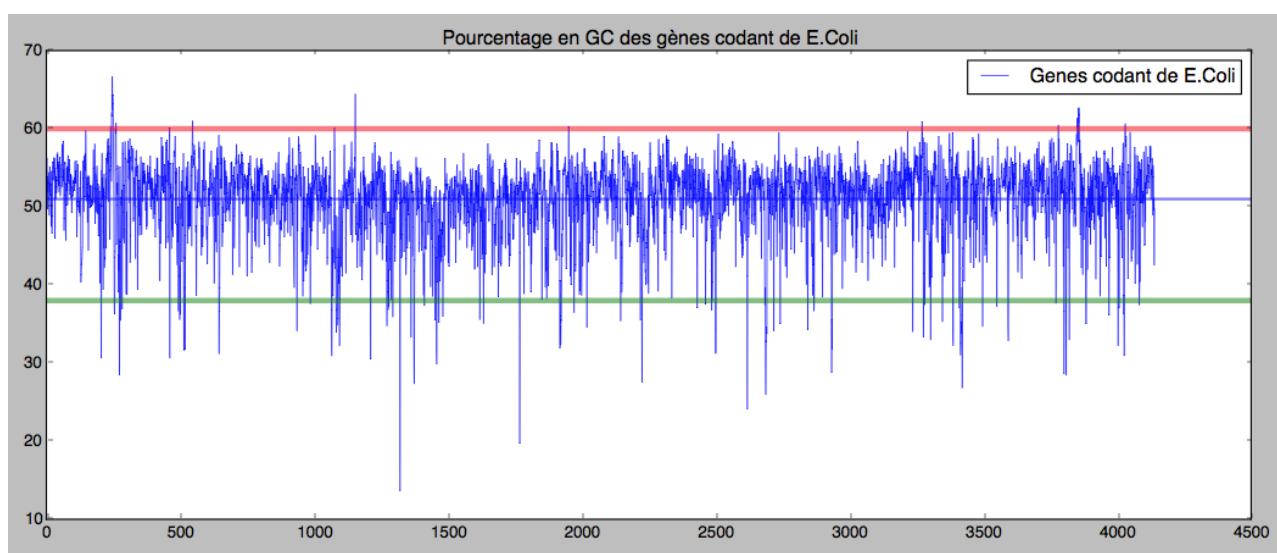
C - Les gènes

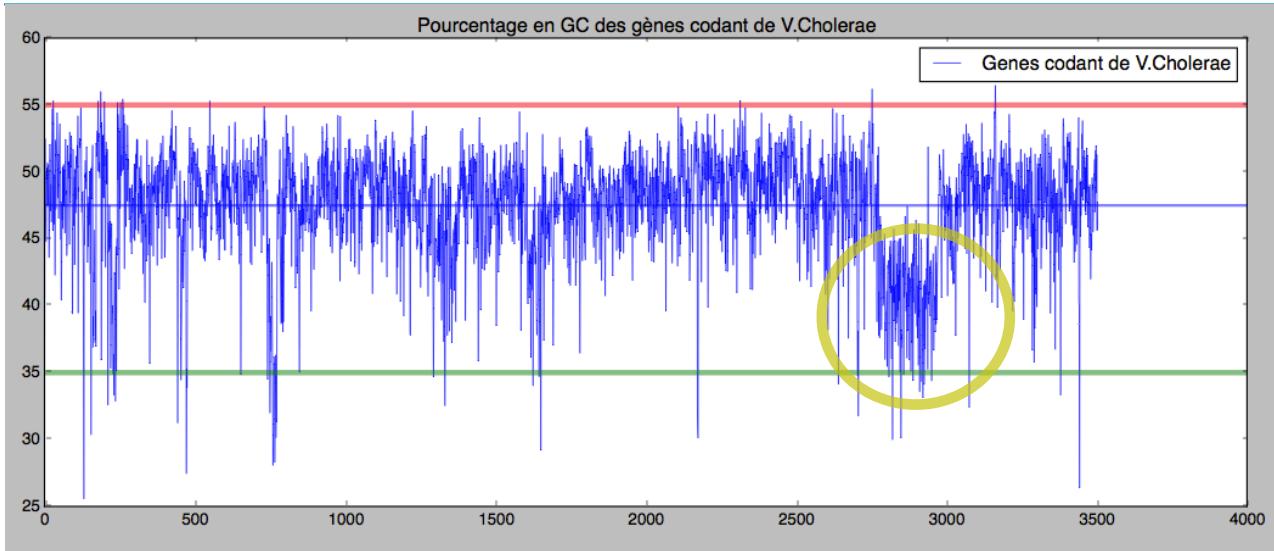
On étudie maintenant les gènes que nous donne le fichier .tab fournit pour le projet.

Au départ nous avons 3503 gènes. Lorsque nous enlevons les gènes ne débutant pas par un Start codon, soit ATG, GTG ou TTG il en reste toujours 3503 . Enfin, si nous filtrons ceux qui possèdent un ou plusieurs stop codons il reste alors 3491 gènes étudiés ici.

Lorsque l'on analyse par bloc on met en évidence les îlots de pathogénicité. C'est différent de l'analyse de la composition des gènes. Il faut alors regarder les le taux de chaque gène pour identifier ceux qui pourraient être issues de transferts horizontaux car trop différents du reste des gènes.

En plus de calculer le pourcentage en G+C pour chaque gène codant de *V.Cholerae* nous avons déterminer la répartition des pourcentages en GC des gènes codants, pour *V.Cholerae* ainsi que pour *E.Coli*.



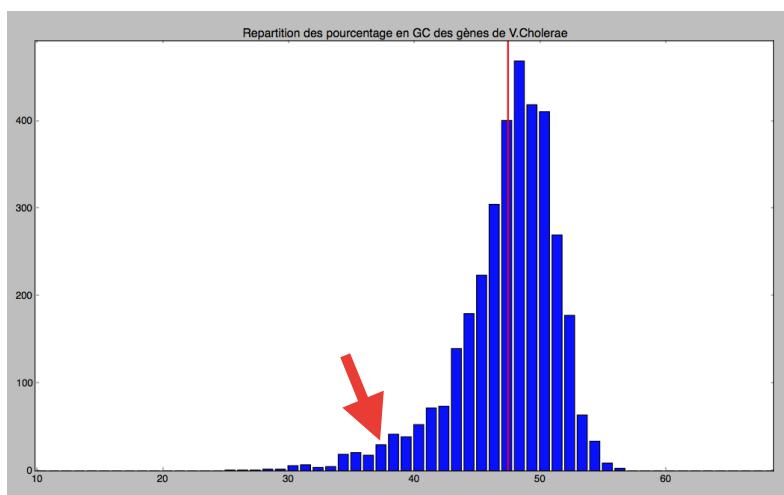


La flèche rouge nous montre une possible anomalie, en effet étant donné la composition globale on remarque qu'une partie s'écarte significativement de la distribution du reste du génome.

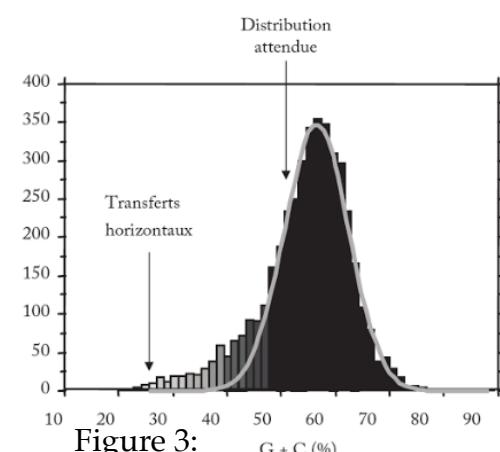
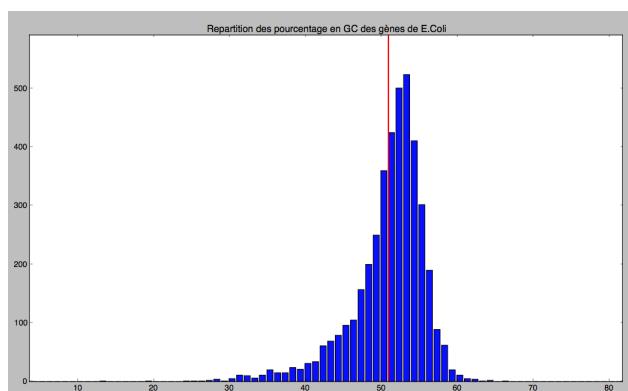
Ils pourraient être interprétés comme le résultat d'un **transfert horizontal** avec un génome riche en A-T.

La ligne rouge verticale représente la moyenne du pourcentage en GC des gènes.

Pour la bactérie *E.coli* qui est l'une des plus étudiées, nous pouvons observer que la distribution attendue, en Figure 3, est similaire



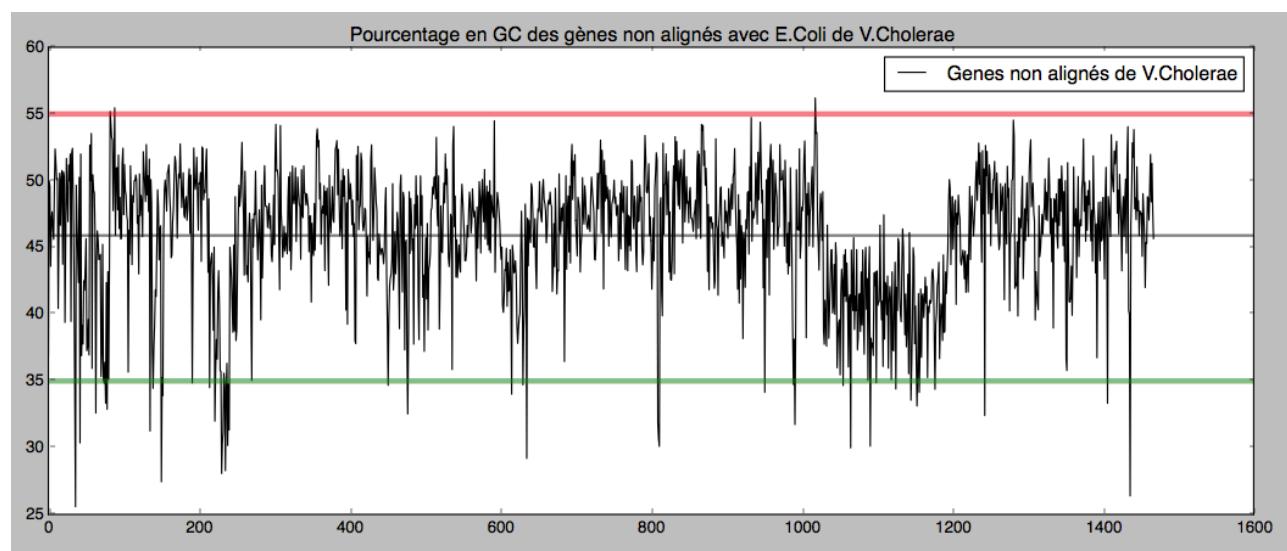
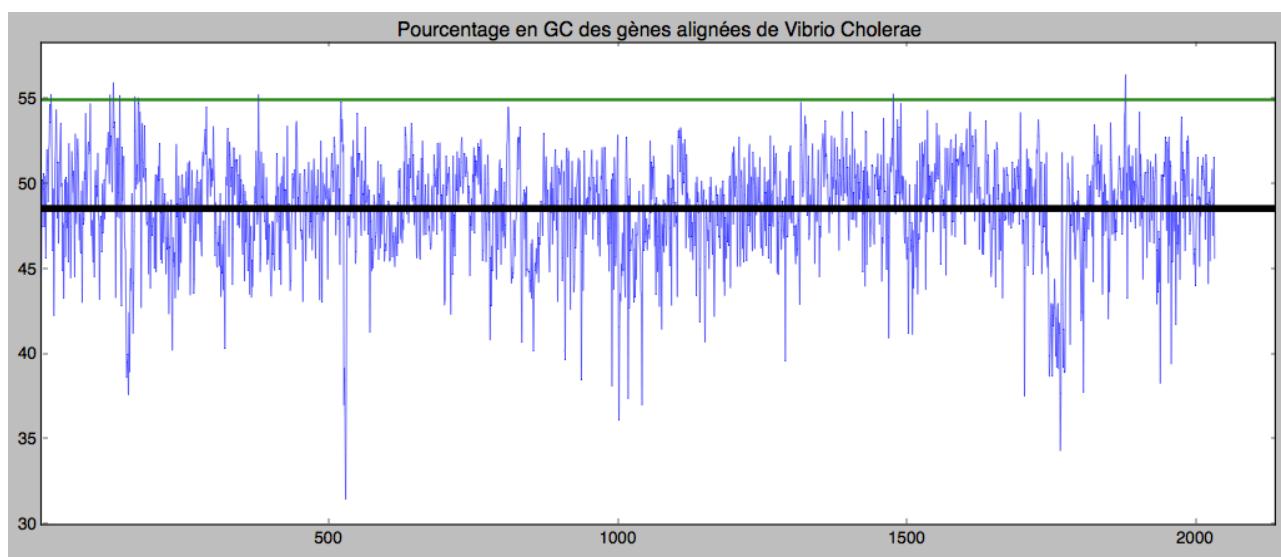
à celle de *V.Cholerae*. Ceci nous permet d'en déduire les transferts horizontaux près de la flèche rouge. Cela peut correspondre aux gènes présents dans le cercle jaune ci dessus.

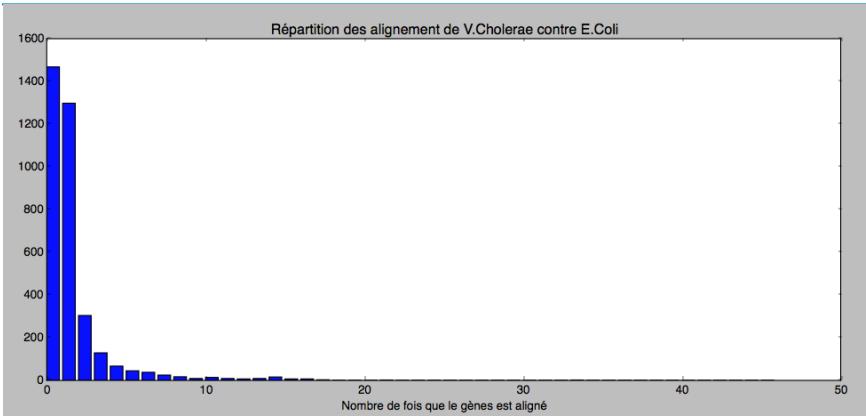


II - Annotation par homologie avec BLAST.

A - Analyse des gènes codants.

On remarque que de nombreux gènes ont plusieurs alignement. En effet, cela peut correspondre à un gène qui a été répliqué plusieurs fois dans le génome. Ce phénomène se nomme la **duplication**. C'est le processus par lequel un gène, un fragment du génome ou le génome entier passe d'une à deux copies, initialement identiques, mais qui peuvent, par la suite, diverger au cours de l'évolution. Le nombre de duplication et leurs longueurs sont très variés. Cependant, les duplications de génome complet sont rares et le plus souvent se sont de petites parties. Pour distinguer les gènes, on dit « orthologues », c'est à dire ceux qui remonte à la divergence entre deux espèces, et les « homologues » c'est à dire ceux qui sont issues de la duplication.





non alignés, soit **45.9%**, donc sensiblement moins que la moyenne global du **génome** de **47,5%** et encore moins que la moyenne des gènes **alignés**, soit **48.6%**. Nous observons la **tendance des gènes non alignés à avoir un taux en GC faible**, contrairement au gènes alignés, du graphe ci dessus.

Etant donné que le pourcentage global en GC de *E.Coli* est de 51% on comprends que ces **gènes alignés établissent plus facilement un lien de filiation entre ces deux bactéries**, en effet les gènes ont un plus faible taux en GC pour *Vibrio Cholerae*. On pourrait déduire que *E.Coli* et *V.Cholerae* ont un ancêtre commun peu éloigné étant donné que les gènes alignés sont plus nombreux que les non alignés mais **qu'ensuite par transfert horizontal la molécule Vibrio à acquis un îlot à caractéristiques différentes**.

B - A propos des non codants.

Reprendons maintenant les gènes non codants, lorsque l'on fait un alignement on voit que les gènes en question n'appartiennent à **aucune famille de gène en particulier**. Aucune valeur n'est alors significative.

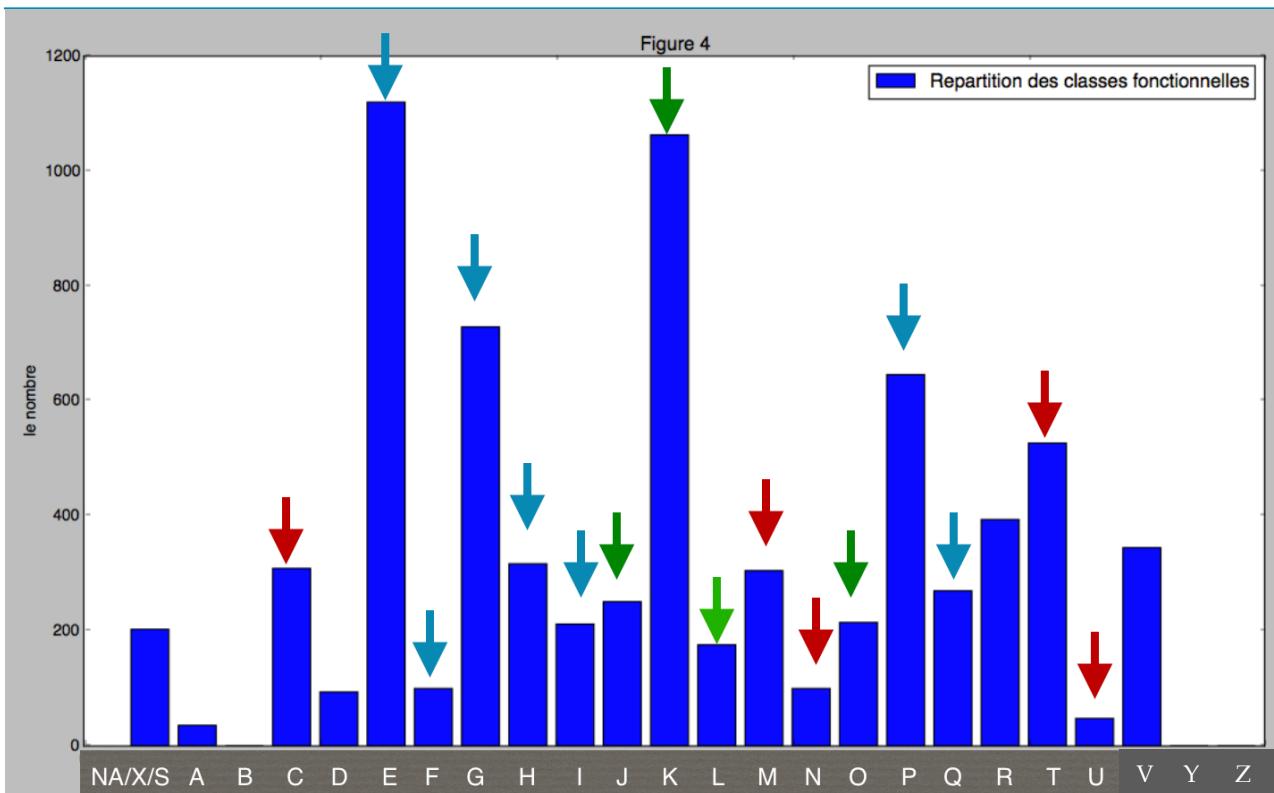
III - Analyse comparative, des propriétés et annotations des gènes correspondants

A- Assignation des catégories fonctionnelles pour les gènes codants avec COG.

Il y a des gènes qui possèdent plusieurs classes fonctionnelles ce qui explique que le nombre de gènes et le nombre de classes représentées ne soit pas égaux. Nous avons choisis de toutes les garder, même pour ceux qui ont plusieurs alignements, afin d'observer la répartition au mieux sans exclure ce qui pourrait être la solution.

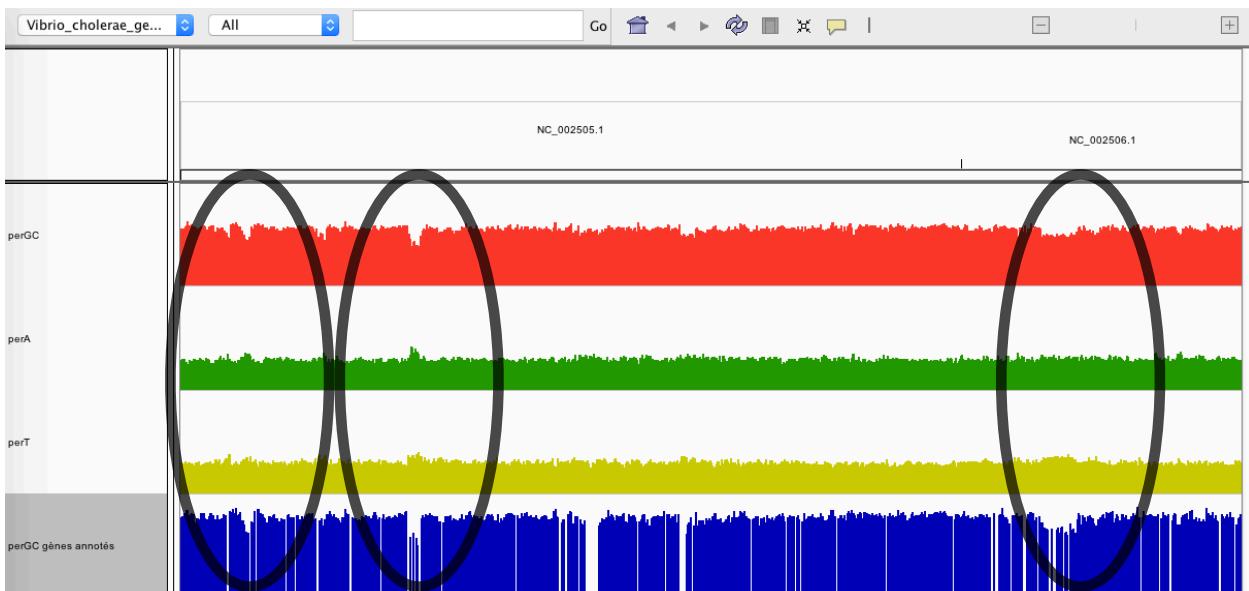
- ↓ Catégorie « Métabolisme et transport » représentant les gènes codant pour des protéines qui soit importent soit conduisent d'autre molécules.
- ↓ Régulation des autres fonction via la communication avec l'extérieur et les liens avec l'environnement.
- ↓ Synthèse des protéines.
- ↓ Synthèse d'ADN.

Pour ce qui est des gènes qui n'ont aucun alignement, ils sont 1468 et le reste de la répartition est dans l'histogramme ci contre. Dans le graphe ci dessus du pourcentage en GC, la ligne noire nous indique la moyenne, en pourcentage GC, des gènes

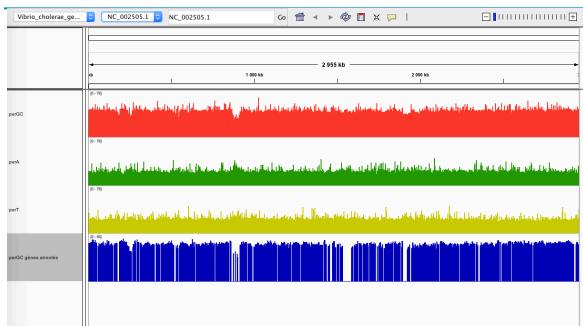


B- Visualisation et navigation dans un génome avec analyse de composition

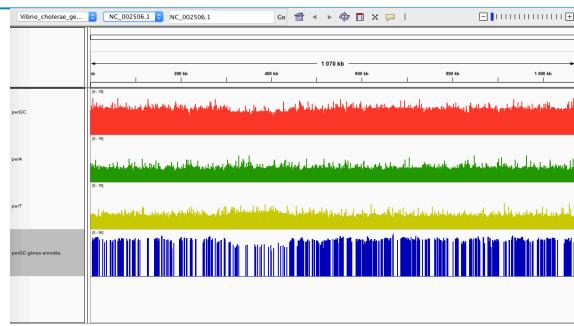
Pour cette figure ci dessous nous utilisons l'outil IGV qui nous permet d'analyser la composition G+C du génome, en rouge, celle en Adénine, en vert, et celle de la Thymine en jaune. En bleu cela correspond à la composition en G+C des gènes annotés. Les cercles noirs représentent les régions qui présente des anomalies par leur



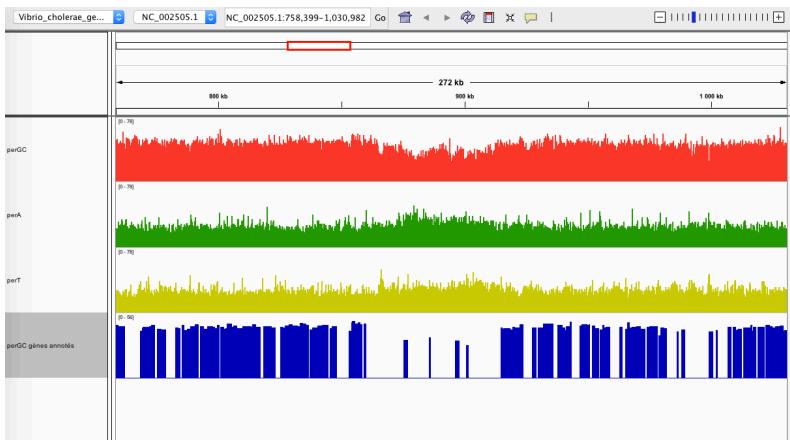
pourcentage en GC.



Chromosome I :



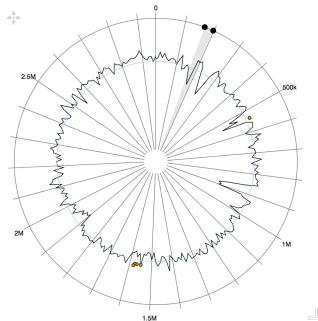
Chromosome II :



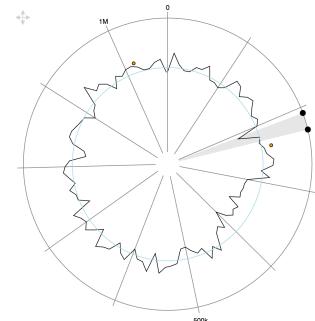
La classe C correspond à des interactions avec l'extérieur ce qui pourrait signifier que si ces gènes correspondent à ceux qui sont pathogènes alors peut être dépendent-il t'interaction avec l'extérieur. La pathogénicité peut alors éventuellement se déclencher par la cause d'éléments extérieurs.

Ci contre nous avons fait un zoom sur les anomalies. Lorsque que regardons les classes fonctionnelle majoritaire des gènes présents dans ces régions atypique nous observons que la majorité est noté C.

Partie B : Annotation plus automatique des îlots de pathogénicité.



Dans cette partie nous utilisons un logiciel, IslandViewer 3, qui nous permet de sélectionner les gènes pathogènes. Nous pouvons les voir sur les figures ci contre prise sur le site pour le chromosome I à gauche, et chromosome II à droite.



I - Les gènes pathogènes via IslandViewer 3.

Les gènes du type PAG ainsi téléchargé sont alors 15, mais seulement 5 différents pour le I et 2 pour le II.

Chromosome I	E.coli	Toutes bactéries	Pourcentage en GC du gène
NP_230156.1	-	-	49.32
NP_231092.1	-	-	41.67
NP_231100.1	-	CTX phage	38.48
YP_007648022.1	-	-	41.27
YP_007648023.1	-	-	41.27
Chromosome II	E.Coli	Toutes bactéries	
NP_232618.1	-	-	47.75
NP_233448.1	-	-	47.08

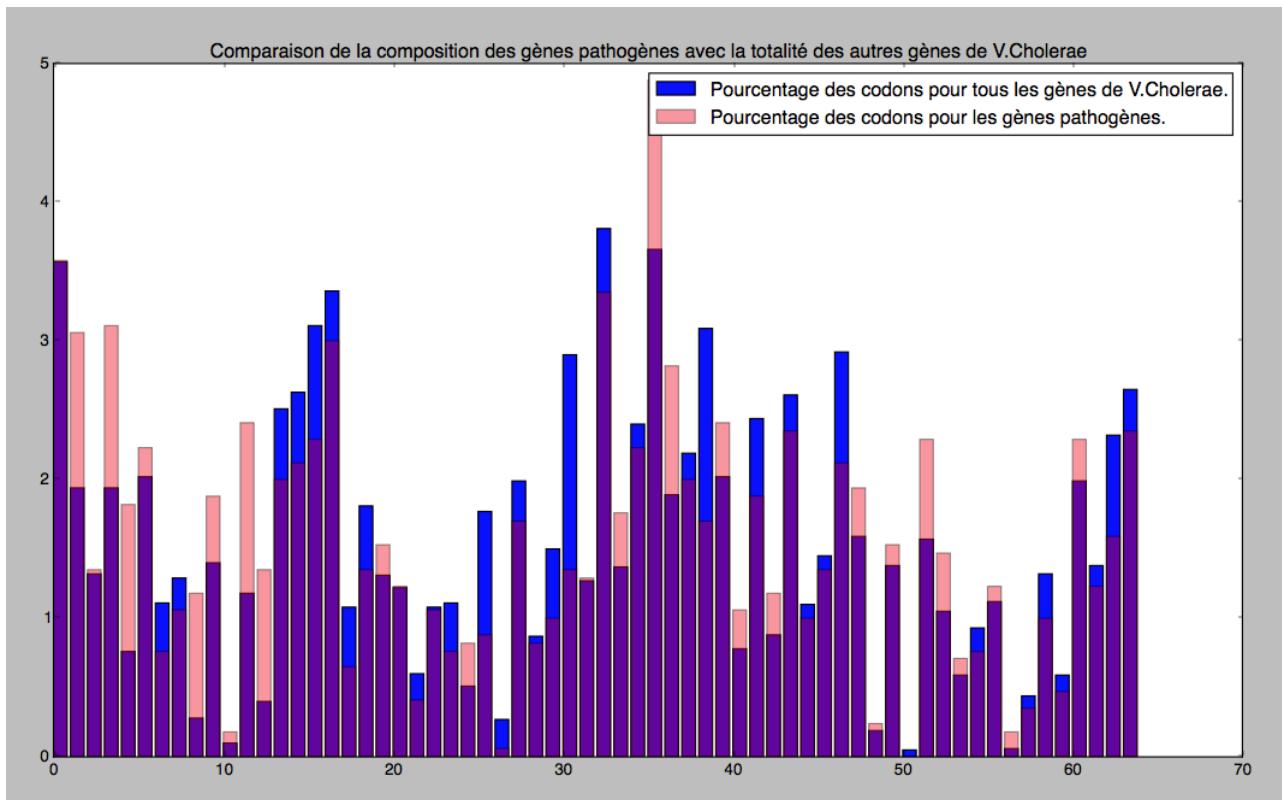
Le gène NP_231100.1 correspond, par BLAST, à une bactérie nommé CTX. C'est un virus de bactérie, c'est à dire « un fragment d'ADN ou d'ARN codant pour une machinerie protéique permettant d'envoyer des copies de lui-même dans d'autres cellules » (*Biologie Evolutive*).

Ce virus rend le *Vibrio* capable de produire une toxine, nommé « choléra enterotoxin subunit A ».

Selon une étude récente du CNRS, CTX est un bactériophage, c'est à dire que pour parasiter le *Vibrio Cholerae* il intègre l'ensemble de son génome dans celui de la bactérie. Ainsi il profite de la multiplication de son hôte afin de se propager. La contre partie est que *Vibrio Cholerae* possède maintenant un gène capable de produire la toxine mortelle

encodée dans le génome du phage. La CTX aurait une facilité à entré dans le génome grâce à l'ilot de pathogénicité présent mis en évidence dans la partie I du projet.

II - Les compositions en codons.



Légende :

0 - AAA	10 - AGG	20 - CCA	30 - CTG	40 - GGA	50 - TAG	60 - TTA
1 - AAC	11 - AGT	21 - CCC	31 - CTT	41 - GGC	51 - TAT	61 - TTC
2 - AAG	12 - ATA	22 - CCG	32 - GAA	42 - GGG	52 - TCA	62 - TTG
3 - AAT	13 - ATC	23 - CCT	33 - GAC	43 - GGT	53 - TCC	63 - TTT
4 - ACA	14 - ATG	24 - CGA	34 - GAG	44 - GTA	54 - TCG	
5 - ACC	15 - ATT	25 - CGC	35 - GAT	45 - GTC	55 - TCT	
6 - ACG	16 - CAA	26 - CGG	36 - GCA	46 - GTG	56 - TGA	
7 - ACT	17 - CAC	27 - CGT	37 - GCC	47 - GTT	57 - TGC	
8 - AGA	18 - CAG	28 - CTA	38 - GCG	48 - TAA	58 - TGG	
9 - AGC	19 - CAT	29 - CTC	39 - GCT	49 - TAC	59 - TGT	

Comme nous comparons les taux en G+C et qu'il diffère en certains, et comme les nucléotides forme les codons qui forment les acides aminé il est fort probable que les gènes à pourcentage en G+C atypique est une **composition en acide aminé atypique**. Dans le graphe ci dessous nous avons choisis d'analyser la composition globale de tous les gènes de *V.Cholerae* (en bleu) ainsi que celle des gènes pathogène (en rouge) donné par le logiciel. Ainsi nous observons que, si pour une majorité des 64 acides aminé, leur

pourcentage de présence est très proche, par exemple AAA, GCA, AGT ... en revanche, pour d'autre tels AAC, AAT, AGC ... ils sont très éloigné. La composition permet donc aussi de mettre en évidence des gènes qui viennent potentiellement d'autre organisme et appuie notre théorie.

Conclusion

La première partie du projet nous a permis de mieux comprendre *Vibrio Cholerae*, d'avoir ces caractéristiques globales afin d'entamer une analyse plus poussée. Dès l'analyse du génome nous avons relevé des irrégularités vers les nucléotides 800 000. Ainsi on peut directement observer ici l'importance que doit avoir le séquençage de génome, et le nombre d'informations que l'on peut déjà en tirer.

Ainsi l'ilot de pathogénicité mis en évidence dans cette partie possède un gène, N16961 nommé tcpA (890449-891123) qui encode pour la toxin co-regulated pilin. Les pili (pilus au singulier) sont des filaments bactériens favorisant l'adhésion de la bactérie aux cellules intestinales.

Aussi l'utilisation en libre services de nombreuses bases de données telles que BLAST ou IslandViewer 3 nous a permis de beaucoup apprendre sur notre bactérie via de multiples comparaisons. Ainsi, les gènes alignés qui ont été mis en évidence contre *E.Coli* permettent d'établir un potentiel ancêtre commun. Mais aussi nous avons déduit des gènes non alignés, aux caractéristiques particulières, qu'ils pouvaient, justement, venir d'une bactérie aux génomes différents.

Enfin la dernière partie a mis la lumière sur les gènes qui sont pathogènes, si la première partie nous a permis de mettre en évidence un îlot de pathogénicité nous ne savions toujours pas et avions aucun moyen de déterminer quel gène était en cause. Ainsi, grâce aux deux logiciels combinés nous avons pu trouver CTX, le coupable de la virulence de *Vibrio Cholerae* O1.

La capacité de certains *Vibrio Cholerae* à être dangereux pour l'homme dépend alors de la présence de ces deux îlots dans leur génome.

En faisant ce projet nous nous sommes rendu compte qu'éventuellement, l'utilisation du taux en G+C pour analyser était un indicateur très efficace mais soulève quelques questions. En effet, cela nous a permis de détecter les transferts horizontaux mais que parce que la composition G + C était très différentes, nous n'aurions donc pas pu mettre en évidence, avec cette méthode, un îlot aux caractéristiques similaires à celle du *Vibrio* étudié ici. Aussi, nous savons que, même si le transfert horizontal a lieu avec une bactérie grandement différente, au cours du temps elle prendra les caractéristiques de son hôte, et sera donc de moins en moins détectable par cette méthode.

Cependant, dans notre cas elle a été suffisante, cela signifie donc que le transfert était avec une espèce éloignée de la sienne mais aussi qu'à l'échelle de l'évolution ce transfert doit être récent.

Sources

Biologie évolutive, Thomas-Lefèvre-Raymond

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC262723/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99527/>

https://en.wikipedia.org/wiki/Vibrio_cholerae

<http://www2.cnrs.fr/presse/communique/737.htm>

<https://www.nature.com/articles/ncomms4549>

<https://sciknowledge.wordpress.com/2013/01/06/genome-evolution/>