UNIVERSITÉ DU QUÉBEC À MONTRÉAL

PROCESSUS DE COALESCENCE POUR UN ÉCHANTILLON STRATIFIÉ

RAPPORT DE STAGE

D'INITIATION À LA RECHERCHE

PAR

CÉDRIC BEAULAC

AOÛT 2011

TABLE DES MATIÈRES

Remerciement

Je tiens d'abord à remercier mon directeur de stage, Fabrice Larribe, pour m'avoir avant tout fait confiance en m'offrant cette chance, puis pour m'avoir soutenue et beaucoup appris tout au long de cet expérience. Je remercie aussi Marie-Hélène Descary ainsi que Mathieu Dupont pour l'aide qu'ils m'ont apporté durant ce stage, ainsi que pour leur support moral tout au long de ces seize semaines.

Bien entendu, je remercie aussi le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour cette bourse de recherche qui m'as permis de rendre cet été constructif, enrichissant et passionnant.

Finalement je tien a remercier famille et amis pour leur soutient tout au long de ce périple.

Merci à tous.

1 Introduction

1.1 Sujet de recherche

Ce travail de recherche vise l'amélioration d'un modèle (Larribe, 2003) de cartographie génétique fine par le graphe de recombinaison ancestral. Lire cette thèse serait important pour bien comprendre l'ensemble de se rapport. De plus, pour bien saisir certains aspects de cette recherche, une base en génétique est nécessaire, base bien détaillée à l'intérieur de cette même thèse.

L'objectif de la cartographie génétique fine est de déterminer la position d'un certain gêne. Dans la plupart des cas, il s'agit de trouver la position d'un gêne muté pouvant causer une maladie. Rapidement, à l'aide d'un échantillon d'haplotypes, nous simulons, dans le passé, la généalogie des ancêtres de ces haplotypes. Nous supposons la position du gêne recherché à plusieurs endroits sur nos haplotypes et tentons de déterminer la position du gêne qui forme les généalogies les plus plausibles.

Comme l'objectif de la méthode de cartographie génétique fine par le graphe de recombinaison ancestral est de déterminer la position précise d'un gêne causal, il est important d'avoir un nombre suffisant de cas à l'intérieur de notre échantillon. C'est pour cette raison qu'un échantillon stratifié est tiré. Néanmoins, le modèle utilisé suppose un échantillon aléatoire simple. L'essentiel de ce travail consiste à trouver comment modifier le modèle actuelle pour prendre en considération ce problème. Le tout consiste d'abord et avant tout, à ajuster les équations de Griffiths (Giffiths, 1997). C'est d'après ces équations qu'est calculé la vraisemblance d'une certaine généalogie. Nous devons modifier ces équations pour prendre en considération le fait que notre échantillon n'est pas réellement aléatoire simple. Pour en arriver au résultat souhaité, nous devrons trouver une manière d'ajuster certain ratios à l'intérieur de ces équations.

1.2 Les équations de Griffiths

Voici l'équation sujette au travail :

Où n_i représente le nombre de séquence de type i dans notre échantillon et n représente le nombre total de séquence dans l'échantillon.

$$
\begin{aligned}
Q(H_\tau) = & \frac{n-1}{n-1+\theta+\rho}\left[\sum_i \frac{n_i-1}{n-1}Q(H_\tau+C_i) + 2\sum_{i\leq j}\frac{n_k+1}{n-1}Q(H_\tau+C_{ij}^k)\right] \\
& + \frac{\theta}{n-1+\theta+\rho}\left[\sum_i\sum_m \frac{1}{L}\frac{n_i+1}{n}Q\left(H_\tau+M_i^j(m)\right) + \frac{nL-a}{nL}Q(H_\tau)\right] \\
& + \frac{\rho}{n-1+\theta+\rho}\left[\sum_i\sum_p \frac{r_p}{r}\frac{(n_j+1)(n_k+1)}{n(n+1)}Q\left(H_\tau+R_i^{jk}(p)\right) + \frac{nr-b}{nr}Q(H_\tau)\right]
\end{aligned}
$$

Dans le modèle avec lequel nous travaillons, trois évènements sont susceptibles de se produire. Soit il y a coalescence de deux séquences, soit il y a une mutation à l'intérieur d'une séquence ou soit il y recombinaison. Ces trois évènements suivent une loi exponentielle avec un taux de $n-1$, θ et ρ respectivement. Ce qui explique le ratio que l'on observe à l'avant des grands crochets quant à savoir lequel évènement se produira en premier.

Le but est donc de trouver un moyen que les proportions des séquences dans l'échantillon à l'intérieur de l'équation soient plus représentatifs de la réalité. Les mutants étants surreprésentés dans notre échantillon de par la manière dont celui-ci a été tiré. Instinctivement, cette problématique peut paraitre simple. Malgré tout, plusieurs étapes sont nécessaires à la résolution de ce problème. Nous devrons d'abord modifier les équations de Griffiths. Puis nous devrons trouver une manière d'estimer la proportion de mutant dans la population au fil des générations jusqu'à l'apparition de cette mutation. Nos recherches ainsi que des séances de discussion nous ont permis d'en tirer

trois méthodes à mettre au point dans le but de résoudre ce problème. Ces méthodes seront présentées dans la prochaine section.

1.3 Modification des équations de Griffiths

Les équations sont modifiés dans le but d'ignorer la méthode d'échantillonnage. Où $X(\tau)$ représente la proportion de mutant dans la population à l'instant τ, ϕ représente l'ensemble des séquences mutantes dans l'échantillon et ω l'ensemble des séquences non-mutantes toujours dans notre échantillon.

$$
\begin{aligned}
Q(H_\tau) = \frac{n-1}{n-1+\theta+\rho}\big[&X(\tau)\sum_{i\in\phi}\frac{n_i-1}{n_\phi-1}Q(H_\tau+C_i)\\
&+(1-X(\tau))\sum_{i\in\omega}\frac{n_i-1}{n_\omega-1}Q(H_\tau+C_i)\\
&+2\big(X(\tau)\sum_{i\le j}\frac{n_k+1-\delta_{ik}-\delta_{jk}}{n_{\in\phi}-1}Q(H_\tau+C_{ij}^k)\\
&+(1-X(\tau))\sum_{i\le j}\frac{n_k+1-\delta_{ik}-\delta_{jk}}{n_{\in\omega}-1}Q(H_\tau+C_{ij}^k)\big)\big]\\
&+\frac{\theta}{n-1+\theta+\rho}\big[X(\tau)\sum_{i\in\phi}\sum_m\frac{1}{L}\frac{n_i+1}{n_\phi}Q\left(H_\tau+M_i^j(m)\right)\\
&+(1-X(\tau))\sum_{i\in\omega}\sum_m\frac{1}{L}\frac{n_i+1}{n_\omega}Q\left(H_\tau+M_i^j(m)\right)+\frac{nL-a}{nL}Q(H_\tau)\big]\\
&+\frac{\rho}{n-1+\theta+\rho}\big[X(\tau)\,(1-X(\tau))\sum_i\sum_p\frac{r_p}{r}\frac{(n_j+1)(n_k+1)}{n_\phi(n_\omega+1)}Q\left(H_\tau+R_i^{jk}(p)\right)\\
&+(1-X(\tau))\,(1-X(\tau))\sum_i\sum_p\frac{r_p}{r}\frac{(n_j+1)(n_k+1)}{n_\omega(n_\omega+1)}Q\left(H_\tau+R_i^{jk}(p)\right)\\
&+\frac{nr-b}{nr}Q(H_\tau)\big]
\end{aligned}
$$

En quelques lignes, ce sont les ratios qui prennent en considération le type de la séquence par rapport au nombre total de séquence dans l'échantillon qui doivent subir des modifications. Pour se faire, nous avons séparé les mutants et les non-mutants en deux groupe distincts. Nous multiplions ce qui concerne les mutants par la proportion

de mutants au même moment dans la population total et le ratio qui en suis est celui du type de séquence par rapport au nombre total de séquence mutante dans l'échantillon. Le même genre de modification a été apporté en ce qui concerne les séquences non-mutantes. Ainsi, nous obtenons une proportion plus réaliste de ce que représente tel type de séquence parmi la population.

2 Méthodes d'estimation de la fréquence de la mutation

Comme énoncé ci-haut. L'une de nos priorités est de trouver un moyen d'estimer la fréquence de la mutation au fil des générations. Voici trois méthodes à l'essais afin de résoudre notre problématique.

2.1 Coalescence de la population proportionnelle à l'échantillon

Ce modèle proposé par Mathieu Dupont est intuitif et simple à utiliser. Bien qu'il ne soit pas très élégant, ni très près de la réalité, sa facilité d'application justifie qu'on l'utilise pour tenter de résoudre le problème.

Dans ce modèle, nous supposons connu la proportion actuelle de mutant de la population total, la taille la population ainsi que la proportion de mutant à l'intérieur de notre échantillon. La particularité de ce modèle est de supposer que que les coalescences des lignées mutantes dans la population sont proportionnelles aux coalescences des mutants à l'intérieure de notre échantillon. Donc, à chaque fois que nous avons un évènement de coalescence de deux mutants dans notre échantillon, nous supposons un nombre de coalescence entre mutants dans la population proportionnellement à ce que représentais les mutants de notre échantillon. Nous allons par la suite recalculer la proportion de mutant dans la population pour ré-ajuster les équations de Griffiths.

Par exemple, supposons une population total de 1000 personnes dans laquelle se trouve un total de 100 mutants. La proportion de mutant dans la population au temps 0 est alors de 10%.Supposons désormais que nous possédons 10 mutants dans notre échantillon. Lors d'une coalescence entre deux mutants dans notre échantillon nous allons admettre que le nombre de mutants dans la population passe de 100

à 90, comme s'il s'était produit 9 autres coalescences en parallèle à celle qui s'est produite dans notre échantillon. La proportion de mutant dans la population, ce que nous cherchons à estimer, sera désormais de 9%.

2.2 Processus de diffusion

Une solution est d'utiliser les travaux de Coop & Griffiths 2004 qui approximent l'évolution de la fréquence du nombre d'allèle mutante à l'aide d'un processus de diffusion.

Le bouquin de Karlin & Taylor 1981 contient un chapitre entier portant uniquement sur le processus de diffusion. Il fût notre principale source d'information à ce sujet. Rapidement, un processus de diffusion est en quelque sorte un mouvement Brownien, le mouvement Brownien étant un cas particulier du processus de diffusion, dans lequel la variance n'est pas nécessairement constante en tout point contrairement au mouvement Brownien. On peut dont imaginer un semblant de mouvement Brownien varier entre 1 et 0 représentant la proportion d'allèles mutantes à travers le temps.

On trouve plusieurs modèles de processus de diffusion à l'intérieur du livre de Karlin & Taylor, tout dépendamment des caractéristiques du modèle de coalescence utilisé. Certains modèles incluent la sélection génétique, d'autre les mutations mais aucun ne prend en considération la recombinaison. Puisque notre modèle ne considère pas la sélection, et que nous suivons la proportion d'une allèle mutante à partir du présent jusqu'à son apparition, le modèle de processus de diffusion qui s'agence le mieux avec notre modèle serait celui sans sélection ni mutation conditionné à l'absorption en 0, moment qui représentera la naissance de la mutation.

2.3 Ajustement du nombre de non-mutant par clonage

Cette méthode, suggérée par F. Larribe, fut l'objet du travail de G. Boucher. Cette méthode peut sembler surprenante mais pourrais très bien s'avérer la plus efficace. En voici les détails.

Comme pour les deux autres techniques, nous devons connaître la proportion actuelle de mutant à l'intérieur de la population. Nous allons ensuite ré-ajuster le ratio mutants contre non-mutants à l'intérieur de notre échantillon en copiant des chaînes d'ADN de non-mutants, choisi aléatoirement parmis ceux de notre échantillon. Nous allons en quelque sortes cloner plusieurs non-mutants jusqu'à l'obtention de la proportion de mutant de la population. Puis, nous allons simplement utiliser la méthode déjà mise en place par la suite sans ajuster les équations de Griffiths.

3 Programmation

Dans cette section, je vais tenter de détailler au maximum les étapes relatifs à l'implantation des diverse méthodes dans le programme actuelle. Sachez que le tout est programmé en C++. De plus, prenez notes que par but de simplification du problème, le modèle de pénétrance utilisé est 0 0 1, ce qui signifie que pour qu'une personne soit considéré comme un cas, c'est que ces deux séquences possèdent le gênes mutant.

3.1 Implantation des nouvelles équations

Il faut désormais implanter le tout dans le programme actuelle mis en place par le professeur Fabrice Larribe, Marie-Hélène Descary et Mathieu Dupont. Tout d'abord, les équations de Griffiths ont dû être modifiées comme indiqué précédemment. Pour se faire, nous devions avant tout implanter une méthode pour calculer le nombre de mutant tout au long de la simulation. Un vecteur fût créé ayant comme seul but d'indiquer si le type de séquence était mutante ou non. Puis, une méthode fût créée pour calculer le nombre de mutant à l'aide de ce même vecteur ainsi que du vecteur contenant la quantité de chacun des haplotypes se trouvant dans notre échantillon. À l'aide du nombre de mutants, nous avons tous le nécessaire pour implanter les équations de Griffiths modifiées telle qu'écrite ci-dessus.

3.2 Implantation de la méthode Dupont-Beaulac

Par la suite nous devrons implanter les méthodes de variation de la fréquence de la mutation à travers le temps. La méthode Dupont est la plus simple a implanter et sera donc idéal dans le but de vérifier que les étapes précédentes ont bien été réalisées. Dans le programme actuelle une méthode s'occupe de faire une mise à jour

de l'échantillon. C'est dans cette méthode que la modification de la fréquence de la population sera effectué. Prenez note que la fréquence de la mutation à l'instant 0 est connue, nous allons seulement la modifier à travers le temps. Nous allons donc utiliser la méthode construite précédemment pour calculer le nombre de mutant avant la mise à jour, puis nous allons calculer le nombre de mutant après la mise à jour. Si le nombre de mutant après est plus petit que le nombre de mutant avant, c'est qu'il y a eu coalescence de deux séquences mutantes. Dans ce cas, nous allons donc modifier la fréquence de mutation de la population proportionnellement à ce que représentait ces deux séquences dans notre échantillon comme expliquer à la section 2.1 .

3.3 Implantation de la méthode par processus de diffusion

Le temps est encore dur à saisir pour l'être humain. Cet été, celui-ce a passé relativement vite et les deux autres méthodes à utiliser n'ont pas pu être implanté. Nous tenterons tout de même de détailler le plus possible comment il serait possible de programmer le tout.

Comme pour l'implantation de la méthode Dupont-Beaulac, la grande partie des modifications se feront à l'intérieur de la mise a jour. Chaque mise à jour sera considéré comme une étape et la fréquence de mutation sera modifié en fonction du processus de diffusion à chacune de ces étapes. Pour en arrivé a des résultats concluants, nous devons nous assurer que la fréquence de mutation se rapproche graduellement de 0, son état absorbant, sans l'atteindre avant la disparition de la mutation dans notre échantillon. Malheureusement, comme expliquer ci-haut, le manque de temps nous a empêché de mettre parfaitement au point cette méthode.

3.4 Implantation de la méthode par clonage

Cette méthode serait relativement simple à implanter. Tout d'abord, nous devons prendre une version original du programme, sans les modifications apportés aux équations de Griffiths, puis nous devons modifier l'échantillon au tout début du programme. Sachant le fréquence de la mutation lors du début de l'exercice le programme devra d'abord tirer aléatoirement, avec remise, une séquence parmi la population de témoins de notre échantillon. Puis la copier et l'ajouter à notre échantillon et ce, jusqu'à ce que notre échantillon aie la même fréquence de mutation que la population. Par la suite, le programme fonctionnera comme avant sans savoir que plusieurs séquences sont en fait des clones de d'autres séquences.

Par contre, dans le cas où la fréquence de mutation est très petit, la taille de notre échantillon après clonage risque d'être très grande et par le fait même le déroulement du programme très long.

4 Résultats

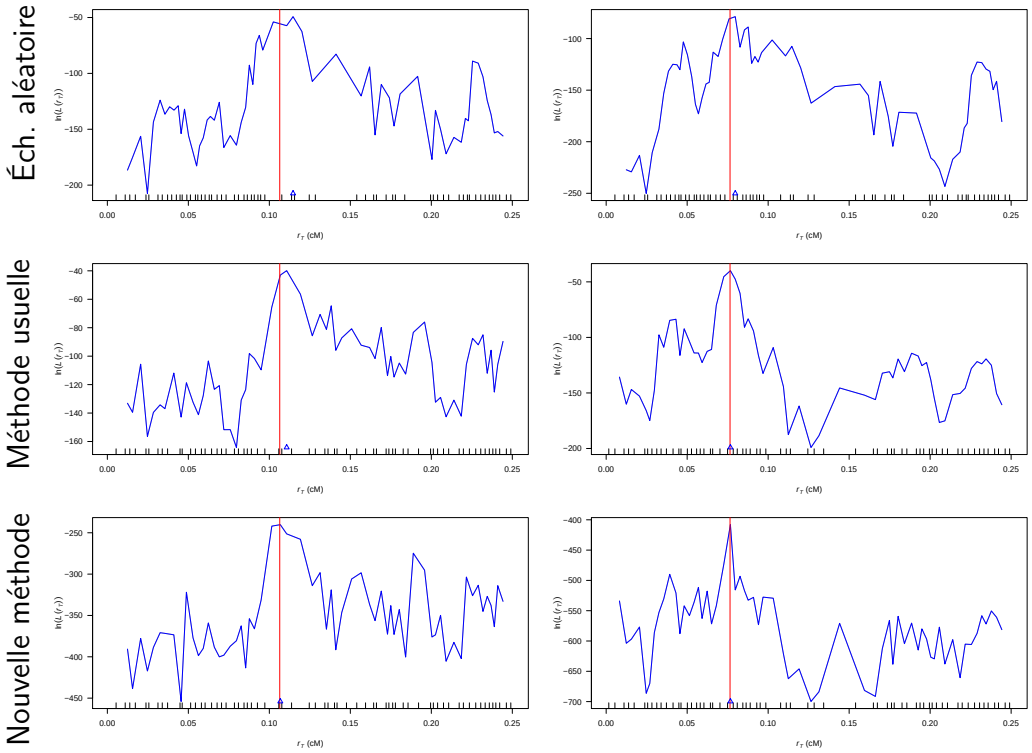
Dans cette section, divers résultats y seront présentés, puis nous tenterons d'analyser ces résultats le mieux possibles pour en tirer d'intéressantes conclusions.

4.1 Graphiques comparatifs

Dans la section ci-dessus y sera présenter les résultats pour trois populations différentes avec pour chacune d'elle 4 fréquences de mutations différentes. Le premier graphique est celui tirer d'une simulation avec un échantillon aléatoire. Il sert quelque peu de référence à quoi devrait plus ressembler la nouvelle méthode. Le deuxième graphique est celui tirer de notre vieille méthode, soit d'utiliser un échantillon stratifié pour avoir un nombre suffisant de cas, mais en supposant un échantillon aléatoire simple. Finalement, le troisième graphique est celui obtenue à partir de notre nouvelle méthode, en utilisant ce même échantillon stratifié mais avec les équations tentant de corriger la supposition d'un échantillon aléatoire simple.

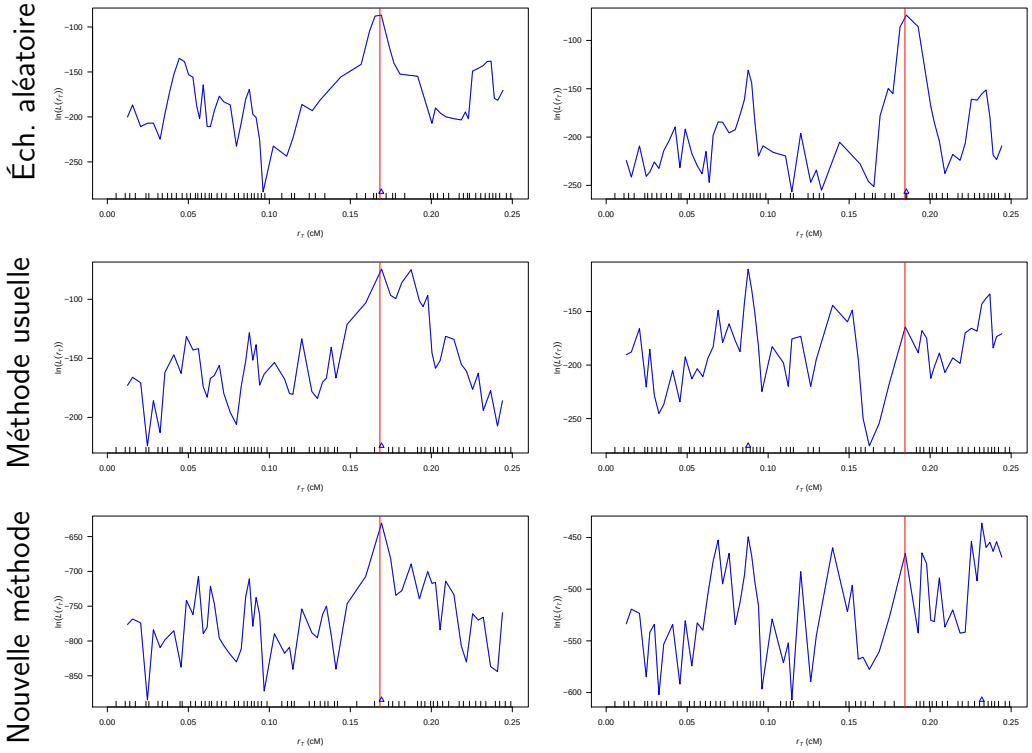
Fréquence du TIM = 0.05 | Fréquence du TIM = 0.10

Éch. aléatoire / Méthode usuelle / Nouvelle méthode



Fréquence du TIM = 0.20 | Fréquence du TIM = 0.50

Éch. aléatoire / Méthode usuelle / Nouvelle méthode



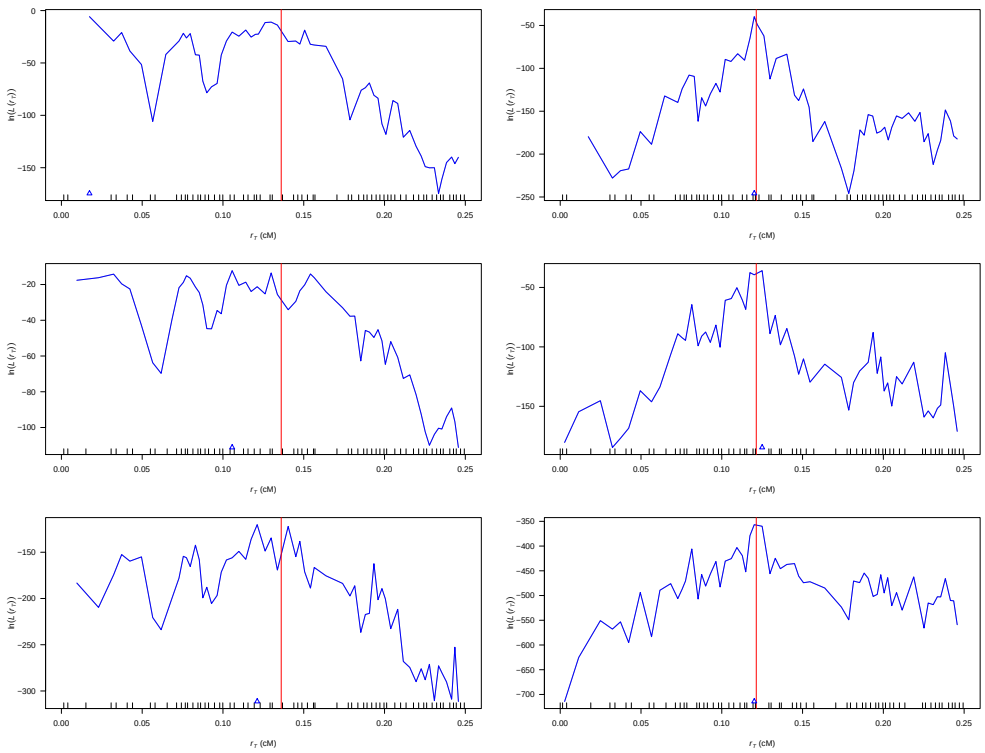
Fréquence du TIM = 0.05

Fréquence du TIM = 0.10

Éch. aléatoire

Méthode usuelle

Nouvelle méthode



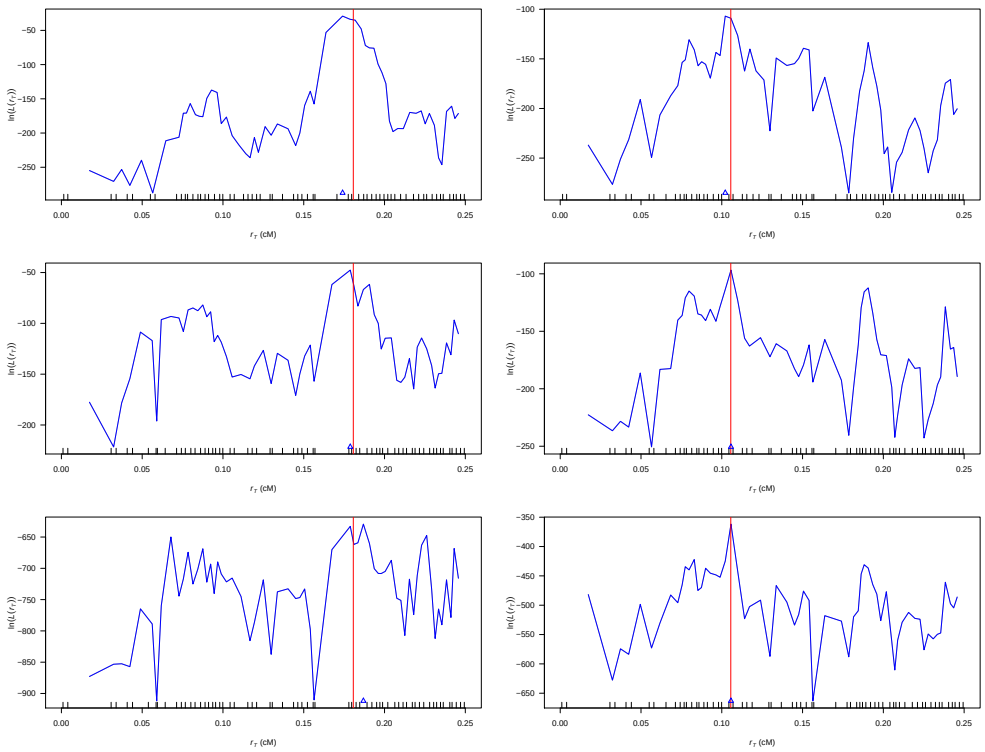
Fréquence du TIM = 0.20

Fréquence du TIM = 0.50

Éch. aléatoire

Méthode usuelle

Nouvelle méthode



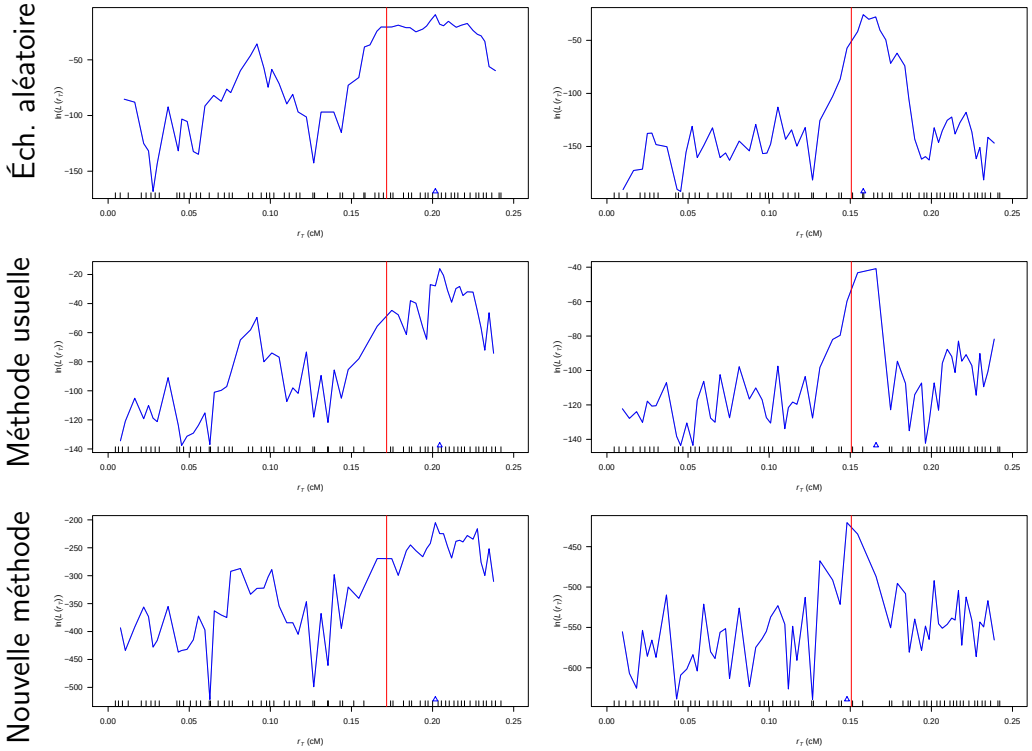
Fréquence du TIM = 0.05

Fréquence du TIM = 0.10

Éch. aléatoire

Méthode usuelle

Nouvelle méthode



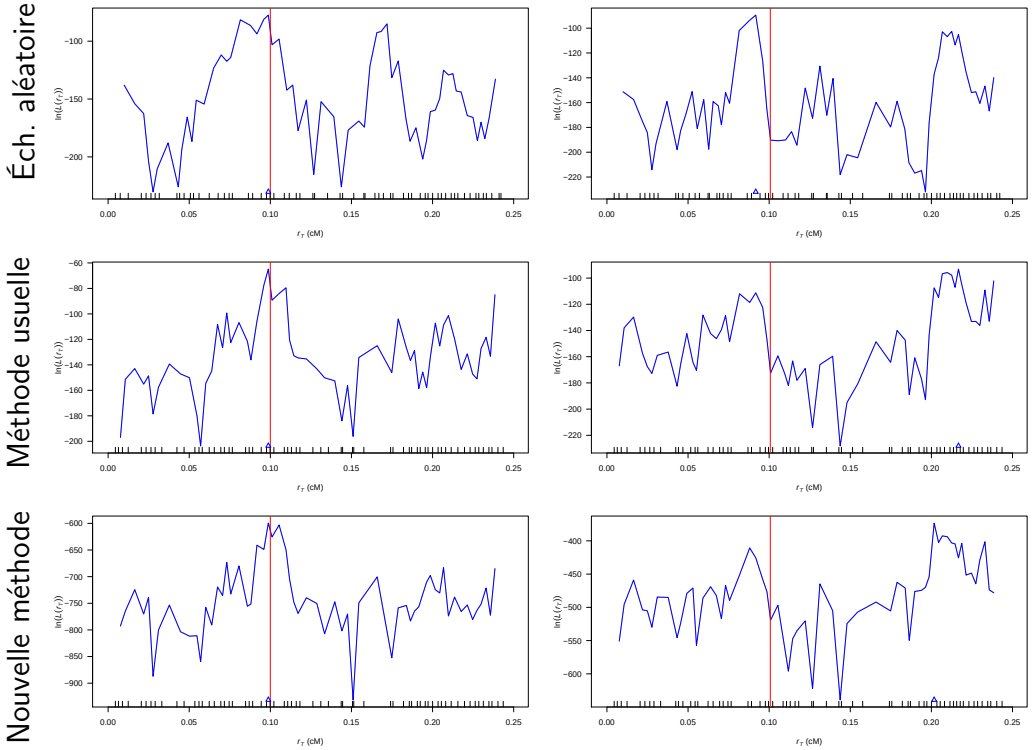
Fréquence du TIM = 0.20

Fréquence du TIM = 0.50

Éch. aléatoire

Méthode usuelle

Nouvelle méthode



4.2 Analyse des résultats

Il n'est pas tâche facile de tirer de grandes conclusions à partir de quelques graphes. Dans les cas où la fréquence de mutation est faible, soit entre 5% et 20% lors de nos tests, l'allure du graphe de notre nouvelle méthode se rapproche un peu plus de la forme du graphe aléatoire que ne le faisait celui généré par la méthode usuelle. De plus, dans certain cas, l'approximation de la position du gêne est plus précise avec cette nouvelle méthode qu'avec les deux autres moyens testés.

Néanmoins, les résultats obtenues nous ont démontrer que dans certains cas le nouveau modèle semble instable avec des fréquences de mutation trop élevés. De nombreuses simulations ont donné de mauvais résultats avec une fréquence de mutation de 50%. Par contre, il est intéressant de remarquer que lorsque la nouvelle méthode générait un résultat erroné, l'ancienne méthode faisait de même la majorité du temps.

Nous avons aussi constaté que la nouvelle méthode est beaucoup plus sensible aux variations, la forme tend plus vers un graphique en dent de scie que l'ancienne méthode. Introduire notre nouvelle méthode au modèle de Fearnhead-Donnelly serait fort intéressant puisque ce dernier est justement beaucoup moins sensible et plus élégant.

Finalement, il est intéressant de constater que notre nouvelle méthode produit des résultats dont la vraisemblance est beaucoup plus petit.

5 Conclusion

En conclusion, la modification du modèle actuelle pour corriger le fait que nous n'utilisons pas un échantillon aléatoire est l'une des nombreuse étapes dans le but de perfectionner le modèle de cartographie génétique fine par le graphe de recombinaison ancestral sur lequel nous travaillons. Par contre, ce rapport ne démontre que le simple travail accompli lors d'un court stage d'initiation à la recherche de seize semaines.

Néanmoins, les résultats obtenus sont encourageant pour l'avenir de cette idée. Il serait très intéressant de vérifier l'application des deux autres méthodes d'estimation de la proportion de la mutations pour comparer l'efficacité de chacune d'elle et en tirer des conclusions plus éclaircis. La méthode Dupont-Beaulac n'étant certainement pas la plus près de la réalité. De plus, il serait encore plus intéressant d'intégrer chacune des méthodes au modèle Fearnhead-Donnelly qui semble s'avérer plus efficace lors des derniers résultats obtenues par l'équipe de recherche.

Il faut aussi considérer que lors de la programmation plusieurs décisions ont été prises de manière à simplifier le problème. Il serait intéressant de revoir certains de ces choix dans le but de s'approcher encore plus de la réalité. Par exemple, le modèle de pénétrance utilisé tout au long de la recherche fut le modèle le plus simple. Il serait enrichissant de faire de nouveaux tests avec un modèle plus réaliste. De plus, toujours dans le but de simplifier la programmation, un gêne non-ancestral a été jugé comme non-mutant systématiquement par le programme. Un tirage proportionnelle à la fréquence de la mutation pourrait être effectuer pour déterminer si un gêne non-ancestral est mutant ou non. De cette façon le programme serait encore un brin plus complexe, mais peut-être plus performant quant aux résultats.

Finalement, le travail accomplis fut très enrichissant pour notre recherche. Les résultats sont variés mais cette modification de la méthode actuelle a un potentiel à ne pas oublier.