essai-format

Elise GUERET

25 Mai 2018

Table of Contents

header includes = Pour avoir des noms automatiques d’éléments en français lors de la création dun document final au format PDF (par exemple le titre de la table des matières)

Insertion de saut de page Word : <https://datascienceplus.com/r-markdown-how-to-insert-page-breaks-in-a-ms-word-document/>

# Page de couverture

* les logos
* intitulé “Mémoire de stage de fin d’études”
* Elise GUERET
* Nom des tuteurs
* adresse de la fac
* adresse du stage
* année universitaire en cours
* titre du mémoire : **Mise en place d’un pipeline bioinformatique pour la recherche de variants à partir de données de séquençage de *Dicentrarchus labrax* et correction de l’annotation du génome de *Dicentrarchus labrax***

# Remerciements

Erick pour tous les conseils et la rédaction du rapport Bruno pour m’avoir confié ce sujet Pierre-Alexandre données de la puce et vcf des variants Madoka pour avoir continuer ce que j’avais commencé niveau manip Flo pour toutes ses pistes concernant python Les stagiaires ISEM et Darwin pour les repas qu’on a partagé L’équipe ISEM pour leur conseil sur le fond et la forme de ma présentation L’équipe Sr2i pour tous les pbs que j’ai eu avec mon ordi L’équipe mbb pour avoir répondu à mes questions cluster et m’avoir permis d’utiliser la big-mem

# Avant-propos

## Table des matières

## Liste des tableaux et des figures

## Liste des abbréviations

# Résumé et mots-clés (en français et en anglais) en 4ème de couverture

# Introduction

Question à se poser avant la rédaction du rapport :  
\* Quels étaient les objectifs de mon travail?  
\* Quel est le bilan de mon travail?  
\* Quelles sont les informations essentielles et les informations secodaires (mais nécessaires) relatives à mon travail?  
\* Comment organiser ces informations pour les rendre compréhensible à un tier?  
Définir le sujet en termes précis et concis, énoncer les objectifs du travail personnel et les moyens mis en oeuvre et présenter le plan adopté pour la suite du sujet d’étude.

Replacer le sujet d’études dans un contexte plus général. Résumer l’état de l’art via une étude des travaux antérieurs.

Décrire les différents projets européens (RobustBass et CRECHE 2016). Dire à quoi ils servent. Comment les données ont été obtenues? Pour CRECHE voir le rapport dans les archives ifremer et pour RobustBass voir les diapos KoM.

Commencer par parler de Dicentrarchus labrax (MORONIDÉ, TÉLÉOSTÉEN), de l’aquaculture et du projet CRECHE 2016. **Papier : Impact of selective breeding on European aquaculture**

Parler des NGS et plus particulièrement du RNA-seq. **papier : Variant callers for Next-Generation Sequencing Data: A Comparison Study** [GATK parralèlisme](https://software.broadinstitute.org/gatk/documentation/article?id=11059)

Parler ensuite des SNPs et des GWAS (définition et existants).

L’aquaculture désigne l’élevage d’organismes aquatiques (d’eau douce ou d’eau salée) destinés à l’usage ou à la consommation humaine, dans des conditions contrôlées. L’aquaculture implique une certaine forme d’intervention dans le processus d’élevage naturel, comme l’empoissonnement régulier, l’alimentation et la protection contre les prédateurs. L’aquaculture est une activité majeure dans plusieurs pays de l’Union Européenne (UE). L’UE a produit en 2015 6,4 millions de tonnes de poissons (pêche et aquaculture confondues). L’aqualculture représentait 1,3 millions de tonnes (19,7%) de cette production alors que la pêche représentait 80,3%. *source : eurostat* Cette activité, tout comme la pêche est encadrée par une politique commune de la pêche (CFP : Common Fisheries Policy). Cette politique commune limite notamment les quotas de pêche. Cette limitation de la pêche permet de sécuriser les stocks de poissons sauvages et de maintenir les ressources disponibles dans les zones de pêches riche. Elle a également contribué au développement de l’élevage d’organismes aquatiques telles que les mollusques, les poissons, les crustacés et les plantes aquatiques. Aujourd’hui l’aquaculture joue un rôle croissant dans la production alimentaire mondiale. Dans le sud de l’Europe différentes espèces de poissons sont élevées à des fins alimentaires telles que le bar européen (*Dicentrarchus labrax*) ou encore le saumon atlantique (*Salmo salar*). L’espèce dont il est question dans cette étude est le bar Européen (*Dicentrarchus labrax*). Il fait partie des Téléostéen et appartient à la famille des Moronidés. Ce poisson marin est très largement répandu le long des côtes de l’Océan atlantique sud, de la Mer Méditerannée, de la Mer Noire, de la Manche et de la Mer du Nord. Il peut également être retrouvé dans les estuaires des fleuves, les lagons et parfois en rivière. Ces conditions de vie font de lui une espèce euryhaline (qui a la capacité à vivre dans des eaux de salinité variable). *papier de Bruno* A l’état sauvage il vit à une profondeur de 100m. Il peut mesuré plus de 70 cm et pesé plus de 6kg. Il se reproduit en hiver et jusqu’au début du printemps. *source : fao*

L’aquaculture de ce moronidé repésente 57478 tonnes soit 4,4% de la production aquacole de l’UE. *source : eurostat* Elle reposait au départ sur la capture de larves et de juvéniles sauvages. Puis elle s’est largement développée suite aux améliorations des techniques de reproduction. Depuis différents programmes de sélection visant à l’amélioration de caractères ont été mis en place. Les traits sélectionnés en premier sont des caractères économique. C’est-à-dire des caractères qui vont permettre de produire plus et plus rapidement. Parmi ces traits, la résistance au jeûne ou le stress sont particulièrement étudiés. Par exemple le projet **Creche 2016** visait à étudier les individus résitant au jeûne et ayant une forte croissance afin de les sélectionner. Un autre projet, intitulé **RobustBass** vise à sélectionner les individus résistant au stress et au maladies. IL existe différente source de stress telle qu’une infection par un pathogène ou un comportement modifié, par exemple lors de leur confinement ou de leur manipulation. Pour connaître le niveau de stress d’un individu, son taux de cortisol est mesuré. Plus un individus sera stressé plus son taux de cortisol sera élevé. Pour étudier la résistance au jeûne, la masse de chaque individus est mesurée. Si il perd du poids alors il est considéré comme résistant au jeûne. En effet, à des fins financières il est important pour les entreprises réalisant de l’aquaculture d’avoir des individus qui grandissent vite sans avoir besoin de trop manger et qui font parti des individus les moins stressés. Ces traits représentent de multiples étapes de sélection qui tendent à l’utilisation d’outils génétiques, génomiques et moléculaires pour améliorer la quantité produite mais aussi pour étudier et maintenir les espèces aquacoles. *QTL for body weight*

Différentes études sur le Bar européen ont abouti à un 1er assemblage du génome, à la production de cartes de liaison génétiques ainsi qu’à des panneaux d’hybrides. Ces différents outils génétiques et génomiques ont pu voir le jour grâce au développement de nouvelle techniques comme le séquençage de nouvelle génération (NGS) à haut-débit qui permet d’étudier soit le génome entier (whole genome sequencing WGS), le transcriptome ( RNA-seq) ou de réaliser un génotypage (RAD-seq ou ddRAD-seq). Le RAD-seq ou ddRAD-seq permet de découvrir un très grand nombre de SNPs par individus et ce, à un coût réaliste. Ces différentes techniques vont permettre d’évaluer plus précisément l’efficacité de la sélection des individus selon leur phénotype (appartenance à une lignée résistante au jeûne, non stressée, de croissance lente, etc.).

Dans cette étude le criblage des SNPs sera effectué à partir de données de séquençage de transcrits (RNA-seq).

# Problématique - Objectifs

Objectifs : Génotyper les individus considérés, trouver les SNPs liés à une lignée et corriger l’annotation Présentation du cahier des charges : correction de l’annotation, mise en place d’un script pour la découverte de variants à partir de RNA-seq et faire des analyses d’association.

# Matériels et méthodes

Mettre en valeur l’acquisition d’une technique, d’un savoir-faire enrichissant mes connaissances initiales. Décrire la méthode ou le logiciel. Pour la correction de l’annotation : Mise en place de scripts Python. Pour le Variant Calling : script en langage wdl (workflow description language) et les suites d’outils Picard et GATK (Genome Analysis ToolKit). Sans oublier les visualisation à l’aide de R. L’analyse d’association PLINK.

## Design de l’étude

Description des données : comment elles ont été obtenues? Description des étapes (annotation de la puce : pourquoi?, GATK pour rechercher les SNPs, snpEff pour annoter la puce et les SNPs, PLINK pour faire calcul LD et Association)

Pour accomplir les objectifs mentionnés ci-dessus, les étapes suivantes ont été réalisées: développer un pipeline pour corriger l’annotation du génome, développer un pipeline pour extraire les génotypes des individus et réaliser les analyses d’association par famille et d’association génotype/phénotype.

Les données de RNA-seq sont issus du Projet CRECHE 2016. Les ARN ont été extraits à partir de foie prélévés sur des individus sacrifiés appartenant aux quatres lignées suivantes J-/J+ (résistant ou non au jeûne) ou C-/C+ (divergence de croissance). Les étapes préliminaires au séquençage ont été réalisées par la plateforme GenSeq. Ces étapes sont : la conservation des échantillons, l’extraction des ARN ainsi que les contrôles qualités des ces ARN. La plateforme GET (Génomique et Transcriptomique) a quant à elle réalisée la création des banques d’ADN complémentaires ainsi que le contrôle qualité des ses banques. Elle a également réalisée le séquençage et les pré-traitement bioinformatiques. L’obtention des reads a été effectué en paired-end 150 sur 5 lignes de séquenceur Illumina HiSeq3000. Parmi les 65 fichiers d’alignement, 24 ont été sélectionnés pour mener cette étude. Ces fichiers proviennent de 23 individus différents (1 individus est en duplica). Chaque lignée est représentée par 6 individus. Les phénotypes disponibles pour ces individus sont : la lignée, le sexe, l’EAi moyen, la classe de l’EAi ainsi que leur comportement.

Le génome de référence de *Dicentrarchus labrax* utilisé das cette étude est le génome qui a été publié. Assemblage jusqu’à des contigs incomplets plus ou moins longs.

L’annotation du génome de référence de *Dicentrarchus labrax* utilisée est une annotation plus complète que celle qui a été publiée puisque des données de transcriptome y ont été ajoutés.

Le transcriptome utilisé est issu du Porojet CRECHE. En effet, en plus des analyses citées précédemment pour ce projet, une analyse d’expression différentielle a été effectuée par l’équipe de la plate-forme MBB. Cette étude a abouti à la création d’un transcriptome à partir des 65 individus pris en compte dans ce projet.

Liste de variants connus chez *Dicentrarchus labrax*, obtenus par le séquençage de génome complet d’une 60aine d’individus avec à chaque fois les parents et un de leur enfants.

Une puce à ADN contenant 57000 SNPs est en cours de fabrication. Pour ces SNPs seule leur position sur le génome est connu. Il serait utile de déterminer les effets de ces SNPs au sein du génome de *Dicentrarchus labrax*.

## Vue générale des pipelines

Que font-ils ? Quels outils ou logiciels ils utilisent?

Les données d’alignement proviennent d’une analyse bioinformatique qui a été réalisée par la plateforme MBB. Cette analyse comprend deux étapes : l’élimination des séquences de mauvaise qualité par TRIMMOMATIC et l’alignement des reads sur le génome de *Dicentrarchus labrax* par l’outils ou le logiciel HISAT2. Une seconde analyse bioinformatique est nécessaire pour effectuer une recherche de variants. Cette analyse consiste à réaliser une recalibration des reads alignés (Picard et GATK) afin de générer des fichiers BAM qui sont le point de départ de cette étude.

Pour étendre et complémenter l’analyse de ces données, un pipeline complémentaire a été programmé. Il utilise les outils suivants :  
\* Snpeff;  
\* VCFtools;  
\* PLINK;  
\* Package R qqman;  
\* Package R ggplot2.

## Description des pipelines

**Penser à impliquer la méthodo c’est-à-dire pour GATK : de comparer les réplicas ainsi que de voir la différence avec des petits bam (peu de reads) et des gros bam (bcp de reads) pour voir si on trouve les mêmes SNPs ou si il y a une différence notable.** A quoi ils servent?

Mettre un schéma général des principales étapes.

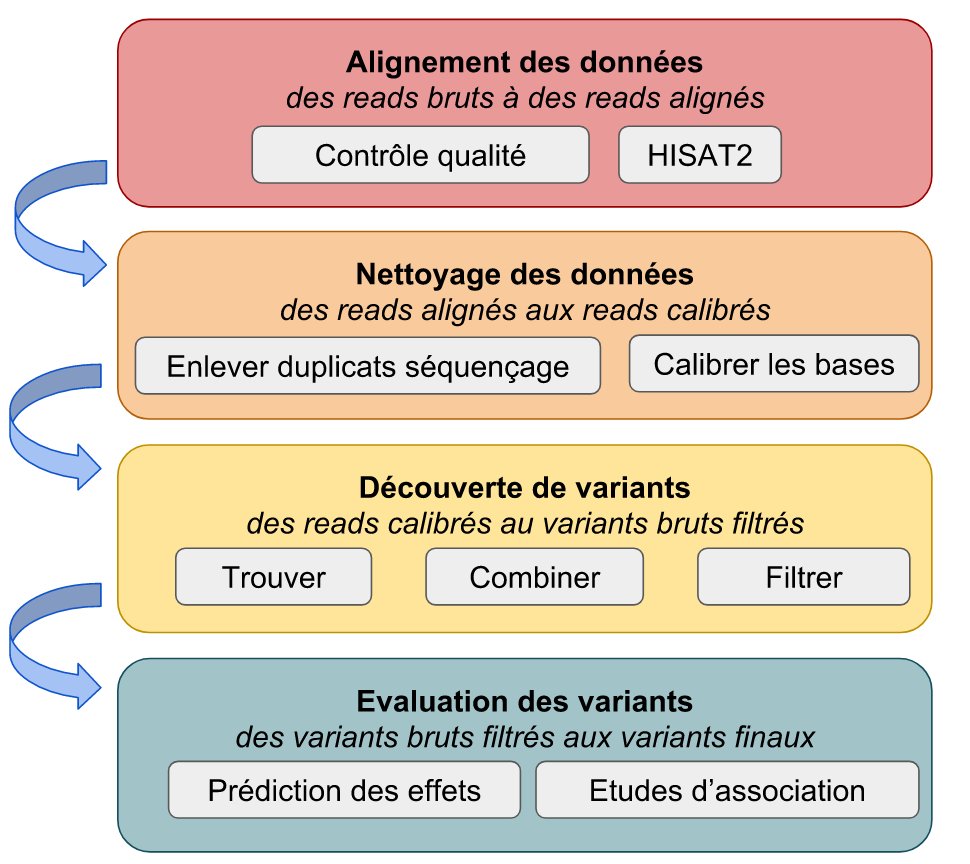
Les avancées technologiques en séquençage à haut-débit et les outils bioinformatiques qui en découlent rendent l’identification et la quantification de variants de plus en plus facile. Cependant, le génotypage de ces variants reste un challenge car chaque outils de recherche de variants est basé sur son propre algorithme de détection de variants. Pour un même jeu de données, il est donc possible d’obtenir des variants différents en fonction de l’algorithme choisi. Il faut donc être conscient que les variants obtenus par un outil peuvent ne pas être retrouvé lors de l’utilisation d’un autre outil. Ceci a été pris en compte lors de l’écriture du pipeline de découverte de variants utilisant GATK, par le respect des [bonnes pratiques de découverte de variants à partir de RNA-seq](https://software.broadinstitute.org/gatk/documentation/article?id=4067) suggérer par l’équipe développant GATK au Broad Institute.

### Correction Annotation par Python

Le pipeline python permet de corriger l’annotation du génome existant. Il utilise des packages biopython. Détails de chaque étape + Schéma

### Pipeline de génotypage

Ce pipeline est décomposé en 4 phases distinctes qui utilisent différentes resources et qui permettent de faire de la parallèlisation pour optimiser les temps de calculs. Ces 4 phases sont les suivantes :  
\* Alignement des données (réalisé par l’équipe de la plateforme MBB);  
\* Nettoyage des données;  
\* Découverte des variants;  
\* Evaluation des variants.

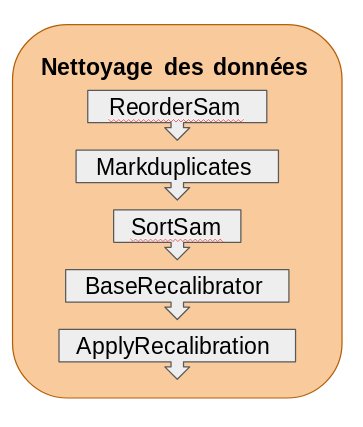


Vue générale du pipeline de génotypage

Ce pipeline est une collection de script wdl, bash et json à exécuter dans l’ordre sur une machine Linux avec une gestion des jobs de type gridengine comme un cluster de calcul. Il est également exécutable sur une machine Linux sans gestion des jobs tels qu’une station de travail.

Pour rendre les données analysables pour la découverte de variants quelques étapes des phases précédentes nécessaires sont détaillées ci-après.

#### Nettoyage des données

Le nettoyage des données est une phase nécessaire pour pouvoir réaliser la suite des analyses. En effet, elle permet d’obtenir des reads calibrés nécessaire à la phase 3. Cette phase comprend des étapes réalisées par Picard ou par GATK. 

La première étape de cette phase appelée “ReorderSam” est réalisée par Picard. Cette étape range les reads en fonction de leur appartenance à un groupe de liaison dans l’ordre des contigs du génome de référence. En effet, si les contigs ne sont pas rangés dans le même ordre cela pose problème à GATK. C’est pour pallier à ce problème que l’on réalise cette étape. Elle est notamment nécessaire lorsque l’alignement a utilisé un ordre différents des groupes de liaisons.

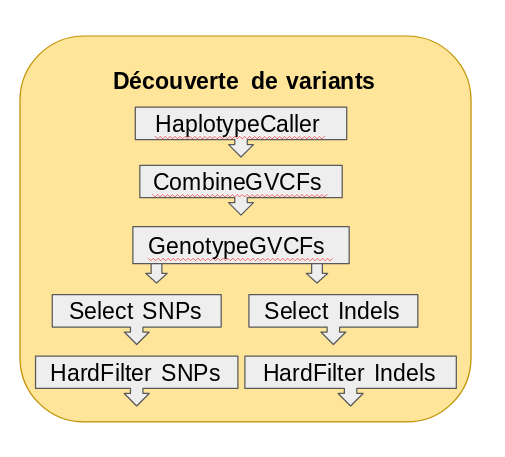
La seconde étape nommée “Markduplicates” utilise Picard. Cette étape est importante car elle marque puis élimine les duplicats de séquençage. Ces duplicas sont des artéfacts dû à la PCR d’enrichissement des banques qui est réalisée en amont du séquençage. Ils sont retirés pour éviter que ces reads soient considérés comme des morceaux de transcrits plus représentés que d’autres.

La troisième étape intitulée “SortSam” est également réalisée par Picard. Cette étape range de nouveau les reads car certains ont été supprimés (masqués) par l’étape précédente. Cela permet d eles mettre en fin de fichier pour ne pas qu’il soit pris en compte dans l’étape qui suit.

Les deux étapes qui vont suivre sont nécessaires car les algorithmes de découverte de variants utilisent la qualité des bases de chaque reads pour déterminer s’il s’agit bel et bien d’une variation ou d’un artéfact. En effet, les scores de qualité sont dépendants du nombre de reads présents. Or certains reads considérés comme des duplicats de séquençage ont été retirés. Il ne faut donc plus les prendre en compte. D’autre part, à ces scores de qualité sont appliqués des erreurs systématiques qui doivent être recalibrés. POur faire cela GATK utilise un algorithmle de “machine-learning” pour modeller empiriquement ces erreurs et ajuster les scores de qualité. Ces deux étapes sont “BaseRecalibrator”et “ApplyBQSR”. La première est réalisée 2 fois afin de visualiser l’efficacité de la recalibration des reads. Cette étape crée une table qui sera utilisée pour la recalibration. La seconde (“ApplyBQSR”) utilise donc la table de recalibration produite par l’étape précédente pour effectuer la recalibration de chacune des bases de chaque reads.

La dernière étape nommée “AnalyzeCovariates” eégalement effectuée par GATK, produit des graphiques montrant l’efficacité de cette recalibration.

#### Découverte de variants

Cette phase est entièrement réalisée par GATK et est la phase clé puisque c’est elle qui extrait les génotypes de chacun des individus. Le succès de cette phase dépend à la fois de la minimisation des Faux positifs et à la fois des Faux négatifs. POur faire cela, GATK procède en plusieurs étapes: la découverte de variants (étape “HaplotypeCaller”) par individus, la fusion des génotypes de chaque individus (étapes “CombineGVCFs” et “GenotypeGVCFs”) par lignée puis la filtration des variants par type de variations (SNPs ou INDELs). Les deux premières étapes ont été designés pour maximiser la sensibilité alors que la filtration permet de maximiser la spécificité par un choix de filtre adaptables à chaque jeu de données. Bien entendu, la découverte de variants dépends du type de données (Whole genome, transcriptome, exome, etc.) mais aussi de d’autres paramètres provenant du séquençage comme la couverture ou la profondeur de séquençage. 

La première étape “HaplotypeCaller” réalise en même temps la recherche de SNPs et d’INDELs par un ré-assemblage *de novo* des haplotypes des séquences actives de la région ciblée. Cet outils est capable de détecter 5 types de variations différentes. [Source1](https://software.broadinstitute.org/gatk/documentation/article?id=3682) Il s’agit des SNPs, MNPs, INDELs, Mixed, et Symbolic. Les SNPs représentent des “single nucleotide polymorphism”. Les MNPs représentent les “multi-nucleotide polymorphism”. Les INDELs représentent des évènements d’insertions ou de délétions de nucléotides. Les Mixed représentent une combinaison de SNPs et d’INDELs à une seule et même position. Les Symbolic montrent qu’il se passe quelque chose à cette position mais qu’on ne sait pas exactement ce qu’il représente. Parfois, HaplotypeCaller utilise l’allèle “non ref” ou \* pour signifier la présence d’une suppression étendue ou des évènements indéfinis comme un trés grand allèle, la perte d’un seul allèle, la perte de deux allèles, etc. [Source 2](https://software.broadinstitute.org/gatk/documentation/article.php?id=6926) HaplotypeCaller produit un fichier au format gVCF (Variant Calling Format) pour chaque individus.

La seconde étape “CombineGVCFs” combine en un seul fichier VCF tous les gVCFs produits dans l’étape précédente.

La troisième étape “GenotypeGVCFs” utilise le fichier VCF de tous les gVCFs combinés de chaque individus pour faire un recalibrage des scores de qualité des variants. Cette analyse réalisée par lignée permet la détection de variants au niveau des locus complexes. Il s’agit d’une aggrégation conjointe multi-échantillons qui fusionnent les enregistrement de manière sophistiquée : pour chaque position du fichier VCF (gVCFs combinés), cet outils combinent tous les enregistrment couvrant le même position, produit des probabilités de génotype correct, re-génotype l’enregistrement nouvellement fusionné puis le ré-annote.

La quatrième étape “SelectSNPs” ou “SelectINDELs” est la sélection des variations selon leur type (SNPs ou INDELs). Cette étape va permettre de séparer les variants afin de leur appliquer des filtres.

La cinquième étape “hardfilterSNPs” ou “hardfilterINDELs” est l’application de filtres.En effet, selon les bonnes pratiques de découverte de Variants à partir de données de RNA-seq, il faut réaliser un “hard-filtering”. C’est-à-dire choisir des valeurs seuils qui définissent quelles données sont considérées comme correcte de celles qui ne le sont pas. Pour cela, ils recommandent différentes valeurs pour différents champs en fonction du type de variation (SNPs ou Indels). Ces valeurs sont présentées dans le tableau suivant :

|  |  |  |
| --- | --- | --- |
| Sigle du champs utilisé | Valeur si SNPs | Valeur si INDELs |
| QD | <2.0 | <2.0 |
| FS | >60.0 | >200.0 |
| SOR | >3.0 | >10.0 |
| ReadPosRankSum | <-8.0 | <20.0 |
| MQ | <40.0 |  |
| MQRankSum | <-12.5 |  |
| InbreedingCoeff |  | <-0.8 |

[Source du tableau](https://software.broadinstitute.org/gatk/documentation/article.php?id=3225)

Ces différents sigles représentent différentes annotation qui signifient :

* **Quality by depth (QD)** : Cette annotation met en perspective le score QUAL du variant en la normalisant par la couverture disponible. Comme chaque reads contribue un peu au score QUAL, les variants dans les régions avec une couverture profonde peuvent avoir des scores QUAL artificiellement gonflés, donnant l’impression que cette variation est soutenue par plus de preuves qu’elle ne l’est réellement. Pour compenser cela, une normalisation de la qualité par la profondeur est effectué, ce qui donne une image plus objective de la qualité de la variation. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_QualByDepth.php)
* **Fischer’s Strand bias (FS)** : Le biais de brin est un type de biais de séquençage dans lequel un brin d’ADN est favorisé par rapport à l’autre, ce qui peut entraîner une évaluation incorrecte de la quantité de preuve observée pour un allèle par rapport à l’autre. L’annotation Fisher’s Strand est l’une des méthodes qui vise à évaluer s’il y a un biais de brin dans les données. Il utilise le test exact de Fisher pour déterminer s’il existe un biais de brin entre les brins sens et anti-sens pour l’allèle de référence ou l’allèle alternatif. La sortie est une p-valeur en Phred. Plus la valeur obtenue est élevée, plus il y a un risque de biais. Et si il y a un biais cela indique la présence de faux positifs. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_FisherStrand.php)
* **Strand Odds Ratio (SOR)** : C’est une autre façon d’estimer le biais de brin en utilisant un test similaire. Le SOR a été créé parce que le FS tend à pénaliser les variations qui se produisent aux extrémités des exons. Les reads issus des extrémités des exons tendent à n’être couverts que par des reads d’une seule direction et le FS donne un mauvais score à ces variations. Le SOR quantt à lui prend en compte les ratios de reads qui couvrent les deux allèles. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_StrandOddsRatio.php)
* **ReadPosRankSum Rank Sum Test for relative positioning of REF versus ALT alleles within reads (ReadPosRankSum)** : Cette annotation teste s’il existe des preuves de biais dans la position des allèles dans les reads qui les supportent, entre les allèles de référence et les allèles alternatifs. Voir un allèle seulement près des extrémités des reads indique une erreur, car c’est là que les séquenceurs ont tendance à faire le plus d’erreurs. Cependant, certaines variations situées près des bords des régions séquencées seront nécessairement couvertes par les extrémités des lectures, donc il n’est pas possible de simplement définir un seuil absolu de «distance minimale à partir de la fin de la lecture». C’est pourquoi un test de somme de rang est utilisé pour évaluer s’il y a une différence dans la façon dont l’allèle de référence et l’allèle alternatif sont supportés. Le résultat idéal est une valeur proche de zéro, ce qui indique qu’il y a peu ou pas de différence dans la localisation des allèles par rapport à la fin des reads. Une valeur négative indique que l’allèle alternatif se trouve aux extrémités des reads plus souvent que l’allèle de référence. Inversement, une valeur positive indique que l’allèle de référence se trouve plus souvent à l’extrémité des reads que l’allèle alternatif. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_ReadPosRankSumTest.php)
* **Root Mean Square of the mapping quality of reads across all samples (MQ)** : Cette annotation fournit une estimation de la qualité de cartographie globale des reads prenant en charge une varaitions. Il produit à la fois des données brutes (somme des carrés et nombre de reads totaux) et le carré moyen calculé. Les données brutes sont utilisées pour calculer avec précision le carré moyen en combinant plusieurs échantillons. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_RMSMappingQuality.php)
* **MappingQualityRankSum Rank Sum Test for mapping qualities of REF versus ALT reads (MQRankSum)**: Cette annotation compare les qualités de cartographie des reads supportant l’allèle de référence avec celles supportant l’allèle alternatif. Le résultat idéal est une valeur proche de zéro, ce qui indique qu’il y a peu ou pas de différence. Une valeur négative indique que les reads supportant l’allèle alternatif ont des scores de qualité de cartographie inférieurs à ceux supportant l’allèle de référence. Inversement, une valeur positive indique que les reads supportant l’allèle alternatif ont des scores de qualité de cartographie plus élevés que ceux supportant l’allèle de référence. Cette annotation peut être utilisée pour évaluer la confiance d’une variation et est une covariable recommandée pour le recalibrage de variante (VQSR). Trouver une différence de qualité statistiquement significative suggère que le processus de séquençage et / ou de cartographie peut avoir été biaisé ou affecté par un artefact. En pratique, seules les valeurs négatives faibles sont filtrées lors de l’évaluation de la qualité des variants car l’idée est de filtrer les variants pour lesquels la qualité des données supportant l’allèle alternatif est relativement faible. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_MappingQualityRankSumTest.php)
* **Inbreeding Coefficient (InbreedingCoeff)**: Cette annotation évalue s’il existe des preuves de consanguinité dans une population. Plus le score est élevé, plus le risque de consanguinité est élevé. Le calcul est une généralisation continue du test de Hardy-Weinberg pour le déséquilibre. [source](https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_annotator_InbreedingCoeff.php)

#### Evaluation des variants et études d’association

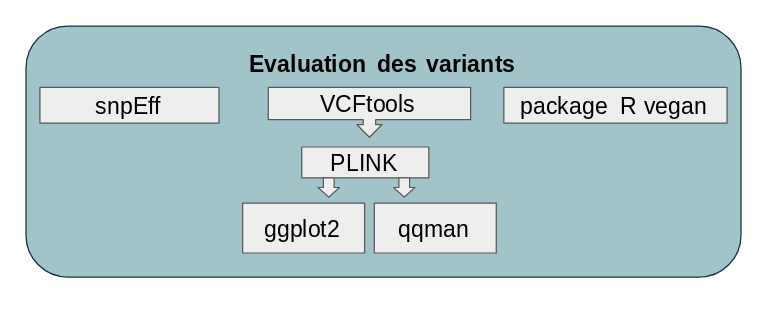
Les phases précédentes ont permis de faire la découverte de variants, d’éliminer les variations les moins soutenues par les données et de produire un fichier VCF. Théoriquement, ce fichier est prêt à être utilisé pour la suite de l’étude. Cependant, il reste à vérifier que ce ne sont pas des artéfacts. Pour cela, il est recommandé d’effectuer des analyses complémentaires pour évaluer la qualité des variants trouvés.

Il faut donc faire la différence entre un jeu de variants considérés comme “bon” d’un jeu de variants considérés comme “mauvais”. Pour cela il existe différente méthodes qui peuvent être appliquer afin de trouver la vérité biologique la plus probable. La méthode la plus fiable est sans doute de faire un séquençage de Sanger des régions proches des variants trouvés. Cependant cette méthode n’est pas envisageable dans cette étude en terme de coût notamment au vu du nombre de variations trouvés. Un autre méthode consiste à évaluer la concordance par rapport aux résultats obtenus à partir d’une puce de génotypage réalisée sur les mêmes échantilllons. Cependant cette méthode ne permet de travailler que sur les variants qui la composent. Mais elle donne tout de même une bonne indication de la sensibilité et de la spécificité. Cette méthode, elle non plus ne peut-être réalisée sur les individus de cette étude car aucune puce de génotypage n’est actuellement sur le marché pour l’espèce *Dicentrarchus labrax*. Une troisième méthode pourrait être de faire une évaluation des variants trouvés en faisant une comparaison de ces variants avec des variants issus d’un autre séquençage sur des individus différents en considérant ce dernier jeu de variants comme étant la vérité. Cependant, cette méthode non plus n’est pas envisagée car aucun jeu de variants issus de séquençage n’est disponible actuellement.

La méthode envisagée ici est donc dans un premier temps d’estimer la qualité globale du jeu de variants obtenus en regardant différentes valeurs. Ces valeurs sont : le nombre de SNPs et d’INDELs obtenus, le ratio Ti/Tv (Transition/Transversion) et le ratio d’INDELs (Insertion/Délétion). Ces valeurs sont calculés par l’outils Picard CollectVAriantsCallingMetrics. IL prend en entrée les fichiers filtrés précédemment ainsi qu’un fichiers de variants connus. Cet outils produit 2 fichiers (Summary et detail) qui sont lus à l’aide du logiciel R. [Source des paragraphes précédents](https://software.broadinstitute.org/gatk/documentation/article.php?id=6308) Ces différentes valeurs sont calculés ainsi:  
+ **Nombre de SNPs et d’INDELs**: CollectVAriantsCallingMetrics recueille le nombre de SNP (polymorphismes mononucléotidiques) et d’indels (insertions et délétions) tels qu’ils se trouvent dans le fichier variants. Il ne compte que les sites bialléliques et filtre les sites multialléliques. De nombreux facteurs influent ces chiffres, notamment la taille de la cohorte, la parenté entre les échantillons, la rigueur du filtrage, l’origine ethnique des échantillons et même l’amélioration de l’algorithme due à la mise à jour du logiciel. Bien que cette métrique soit insuffisante pour évaluer les variants trouvés, elle fournit une bonne base de référence.  
+ **Indel Ratio**: Le ratio d’indel est déterminé comme étant le nombre total d’insertions divisé par le nombre total de délétions; cet outil n’inclut pas les variants filtrés dans ce calcul. Habituellement, le rapport indel est d’environ 1, car les insertions se produisent généralement aussi souvent que les délétions.  
+ **Rapport TiTv**: Cette métrique est le rapport entre les mutations de transition (Ti) et de transversion (Tv). Pour les données de séquençage du génome entier, TiTv devrait être ~ 2.0-2.1, tandis que les données de séquençage de l’exome entier auront un rapport TiTv de ~ 3.0-3.31. Dans le cas de cette étude il s’agit de données de transcriptome, ce type de données étant proche de données d’exome le rapport attendu sera donc aux alentour de 3.0-3.31. [Source des 3 paragraphes ci-dessus](https://gatkforums.broadinstitute.org/gatk/discussion/6186/howto-evaluate-a-callset-with-collectvariantcallingmetrics#latest)

Dans un second temps une annotation fonctionnelle des variations trouvés est réalisée ainsi que des études d’association (famille et génotype/phénotype).

##### Design

Cette quatrième et dernière phase permet donc d’évaluer les variants en terme de qualité, d’effet prédit, d’association par famille ou par phénotype/génotype. Cette dernière phase fait appel à différents outils et logiciels : Picard pour obtenir les différentes valeurs mentionnées ci-dessus, SnpEff pour prédire les effets des SNPs trouvés, PLINK pour réaliser l’étude d’association par famille, le package R vegan pour réaliser l’analyse d’association génotype/phénotype, etc. 

La dernière étape de GATK (“hardfiltering”) produit deux VCFs (un par type de variations : SNPs et INDELs) qui sont les fihciers d’entrée pour les différents logiciels ou outils mentionnés précédemment.

La première étape consiste donc à obtenir le nombre de SNPs et d’INDELs obtenus, le ratio Ti/Tv (Transition/Transversion) et le ratio d’INDELs (Insertion/Délétion) via l’utilisation de CollectVAriantsCallingMetrics (Picard) puis d’analyser ces données à l’aide du logiciel R et son interface graphique RStudio.

La seconde étape consiste à réaliser une annotation fonctionnelle des SNPs potentiels trouvés par la prédiction des effets. Cette étape est réalisée par le logiciel snpEff [aHoff2015] qui analyse le fichier d’entrée, annote les variants et calcule les effets qu’ils produisent sur les gènes connus.

La troisième étape consiste à une analyse d’association par famille. Cette étape est consitituée de sous-étapes. En effet, PLINK a besoin de fichier au format ped et au format map pour procéder aux associations. Or, GATK produit des fichiers au format VCF. L’outils VCFtools est donc nécessaire pour produire les fichiers ped et map à partir de fichiers VCF.

La quatrième étape consiste à une analyse d’association génotype/phénotype.

##### Population stratification

Expliquer et définir les problèmes de stratification de la population.

# Résultats - Discussion

Avoir un fil conducteur dans un ordre logique pour mettre en évidence un point précis ou un but recherché.

Discussion :  
\* Mes résultats sont-ils pertinents?  
\* Quelle est leur signification?  
\* Quelle en est la portée?  
\* Peuvent-ils être utile à d’autres?

## Résultats du génotypage

Parler de Picard et de GATK

Une étude de 2013 a permis de comparer différents outils de découverte de variants. Les outils testés étaient GATK, SAMtools, glftools et Atlas2. Ces différents outils ont été testés sur des données de séquençage d’exome en mode multi-échantillon et en mode échantillon par échantillon. Les auteurs concluent que GATK a le taux de redécouverte (0.9969) et la spécificité (0.99996) les plus élevés. D’autre part son ratio transition-transversion (Ti/Tv) proche de la valeur attendue (3.02 pour des données d’exome). Suite à cette étude, la stratégie GATK a donc été choisie.

## Résultats de l’étude d’association

### Fréquence allèlique et données manquantes

Dans le but d’obtenir les fréquences allèliques ainsi que les données manquantes, PLINK a été utilisé.

### Estimation de l’“ancestry”

### Population stratification

### Top SNPs

### Evaluation des top SNPs

### Analyse fonctionnelle des top SNPs

# Conclusion

Résume le travail accompli et fait apparaître si les objectifs ont été atteints

S’achève par des perspectives ou sur un bilan personnel

* Que dois-je retenir de ce travail?

etblahblah Pour la rédaction de doc en Rmd : <https://stt4230.rbind.io/communication_resultats/redaction_r_markdown/#creation-dun-document-r-markdown-en-rsudio>

# Options de blocs de code R en R Markdown :

Les options de blocs de code R les plus utiles sont les suivantes :

* eval (TRUE par défaut, ou FALSE) : détermine si le code R doit être évalué ou non,
* echo (TRUE par défaut, ou FALSE) : détermine si le code R doit être affiché ou non,
* results ('markup' par défaut, ou 'hide' ou 'hold' ou 'asis') : détermine comment les sorties doivent être affichées,
* error (FALSE par défaut, ou TRUE) : détermine si les messages d???erreur doivent être affichés.
* warning (TRUE par défaut, ou FALSE) : détermine si les messages d???avertissement doivent être affichés.

# References