INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

UNIVERSITÉ DE LYON

ENS DE LYON

Rapport de stage 3A

Identification de mutations affectant la variabilité d'expression par BSA-Seq

Tuteur de stage : Fabien Duveau

Enseignant référent à l'INSA LYON : Samuel Bernard

Période de stage : du 18/06/2020 au 10/07/2020 et du 17/08/2020 au 11/09/2020

Rédigé par :

Julie KLEINE-SCHULTJANN

3ème année Bioinformatique et Modélisation

Département Biosciences INSA LYON

2019 - 2020

Table des matières

Préambule

Le Laboratoire de Biologie et Modélisation de la Cellule a pour but de mieux comprendre le fonctionnement normal et pathologique des cellules à une échelle moléculaire. Les différentes équipes travaillent sur plusieurs phases de la vie cellulaire telles que la division, la prolifération et différenciation, l'intégration tissulaire, la régulation de l'expression génique, la structure et interaction des biomolécules.

Les recherches sont divisées selon 3 axes principaux :

→ Approches quantitatives de la dynamique du génome et de son expression

→ Processus pathologiques : aspects cellulaires et moléculaires

→ Biologie des systèmes : signalisation et processus cellulaires développementaux

Le LBMC regroupe 100 chercheurs, ingénieurs et postdoctorants. Parmi les 15 équipes de recherche, j'ai travaillé sous la tutelle de Fabien Duveau avec l'équipe de Gaël Yvert, responsable de l'équipe Génétique des Variations Intra-Espèce. Ils étudient les mécanismes génétiques fondamentaux responsables des différences phénotypiques entre des individus de la même espèce. Pour comprendre pourquoi certains individus ont une prédisposition à développer certains phénotypes, ils travaillent sur des modèles expérimentaux tels que la levure *S. cerevisiae* et utilisent des outils génomiques et bioinformatiques. J'ai également travaillé avec Laurent Modolo, responsable du pôle bioinformatique, sur toutes les questions d'ordre informatique. Le LBMC travaillant en collaboration avec d'autres laboratoires, j'ai pu analyser des données de séquençage provenant de l'Université du Michigan.

1 Introduction

Les différences phénotypiques entre des individus d'une même population peuvent être dues à des mutations. On observe également des différences phénotypiques entre des individus génétiquement identiques. Ces différences peuvent s'expliquer par des facteurs environnementaux ou des facteurs intrinsèques à la cellule. Par exemple, l'expression d'un gène dépend de la fréquence de liaison du facteur de transcription au promoteur et de la quantité d'ARN messagers transcrits [3]. L'expression d'un gène est donc variable entre les cellules appartenant une même souche. Par exemple, Schaffer *et al.* ont observé que lors du traitement d'un mélanome, la plupart des cellules cancéreuses étaient éradiquées mais que certaines cellules étaient résistantes au traitement [6]. Ces cellules étant génétiquement identiques, un autre mécanisme que les mutations explique leur résistance. Ils ont montré que les cellules cancéreuses résistantes au traitement ont une variabilité d'expression extrême de certains gènes comparée à la variabilité des cellules saines [6].

D'autres études s'intéressent à la variabilité d'expression entre individus isogéniques. Fabien Duveau et ses collègues du Michigan ont identifié les mutations associées à une différence de niveau moyen d'expression du gène TDH3 chez les levures *S. cerevisiae* en utilisant la technique de Bulk Segregant Analysis (BSA) associée à du séquençage nouvelle génération [1]. Cette technique a démontré qu'il était possible de détecter des mutations de type polymorphismes nucléotidiques simples (SNP) ayant de faibles effets phénotypiques.

Ce projet a pour but de déterminer les mutations associées à une différence en variabilité d'expression en fluorescence du gène TDH3 par BSA-Seq. Plus explicitement, la souche sauvage et la souche mutante ont chacune une variabilité d'expression qui leur est propre et on suppose qu'une ou plusieurs mutations affectent la différence de variabilité d'expression entre les deux souches. Tout l'enjeu consiste à détecter ces mutations.

Lors de mon stage, j'ai étudié 6 mutants obtenus par mutagénèse à l'EMS. Ces mutants ont tous une moyenne d'expression significativement identique à celle de la souche sauvage. 3 mutants ont une variabilité d'expression plus faible que la souche sauvage et 3 ont une variabilité d'expression plus forte que la souche sauvage.

2 Méthodes et outils bioinformatiques

2.1 Démarche expérimentale

Cette section explique le protocole suivi par Fabien Duveau lors de son postdoctorat à l'Université du Michigan permettant d'obtenir les données de BSA-Seq que j'ai analysé lors de ce stage.

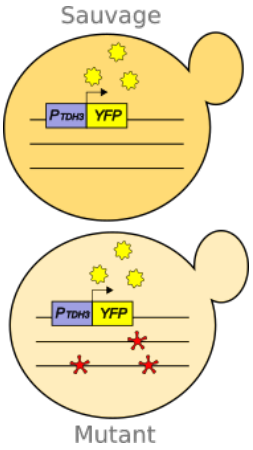


FIGURE 1 – Schéma d'une cellule de levure sauvage et d'une cellule de levure mutée (en rouge les mutations)

Pour déterminer les polymorphismes nucléotidiques simples (SNP) associées à la variabilité d'expression de fluorescence, des cellules de levure *S. cerevisiae* exprimant le rapporteur fluorescent YFP sous contrôle du promoteur du gène TDH3 ont été soumises à une mutagénèse au méthanesulfonate d'éthyle (EMS) (voir Figure 1). Le gène TDH3 est fortement exprimé chez la levure et est responsable de la glycolyse. La variabilité d'expression de ce gène est facilement mesurable grâce au rapporteur fluorescent YFP. Un cytomètre est un appareil à un haut débit faisant défiler une à une les cellules de l'échantillon devant un faisceau laser et mesurant la lumière réémise par les cellules. Il permet de les compter et de caractériser leur taille et leur niveau de fluorescence. Grâce à cet appareil, la distribution de niveau de fluorescence à partir de ≈ 5000 cellules sauvages est connue. De même, la distribution de niveau de fluorescence est connue pour les 1922 mutants étudiés par Metzger *et al.* [5]. Chaque mutant a été cultivé de façon clonale pour obtenir un grand nombre de cellules isogéniques.

Par cytométrie en flux, on obtient la distribution de niveau de fluorescence de cellules isogéniques cultivées dans un même environnement. La Figure 2 représente ces étapes et distributions schématiquement.

Parmi les 1922 mutants sélectionnés et mis en culture par Metzger *et al.* [5], Fabien Duveau en a sélectionné 6 pour ce projet. On appelera ces mutants "mutants EMS" car obtenus par mutagénèse à l'EMS. La Figure 3 représente les 1922 mutants selon leur moyenne et variabilité d'expression. Les 6 mutants EMS sélectionnés (entourés en rouge) ont une moyenne d'expression significativement identique au sauvage. 3 mutants EMS ont une forte variabilité comparée à celle du sauvage et 3 mutants EMS ont une faible variabilité comparée à celle du sauvage. La variabilité correspond à l'écart-type divisé par

la moyenne de fluorescence observée entre cellules isogéniques.

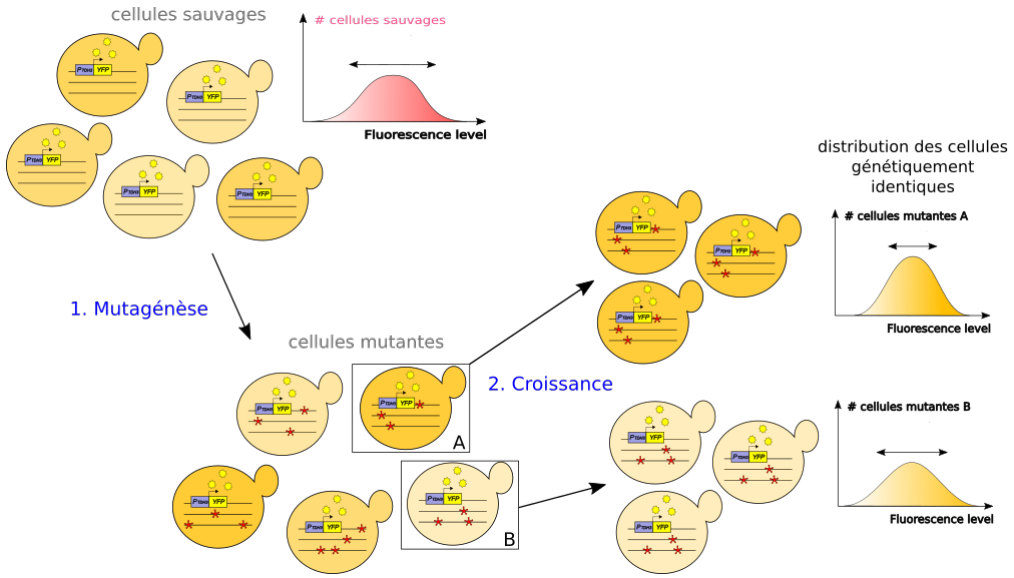


FIGURE 2 – Mutagénèse à l'EMS des levures *S. cerevisiae* puis croissance de certains mutants.

Les cellules sauvages sont soumis à une mutagénèse à l'EMS et les cellules mutantes résultantes sont mises en culture. La cytométrie en flux des cellules sauvages et mutées renseigne sur la moyenne et la variabilité d'expression en fluorescence des cellules.

Le but de l'approche de BSA-Seq est de déterminer quelle(s) mutation(s) sont associées aux différences de variabilité parmi toutes les mutations présentes dans les mutants EMS.

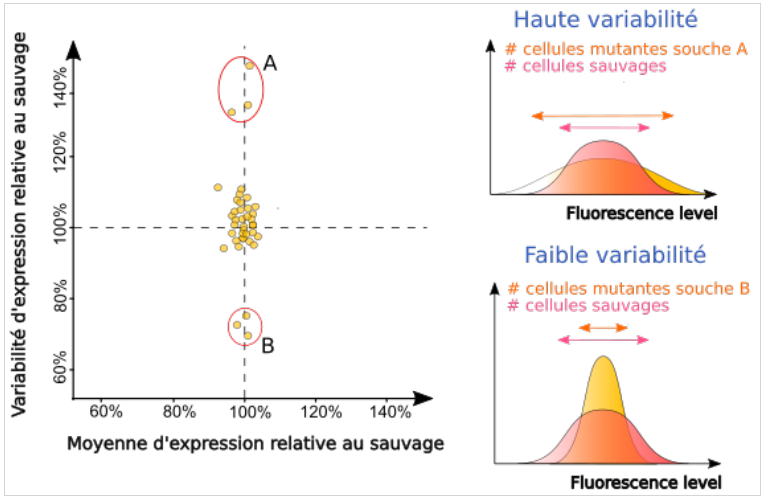


FIGURE 3 – Sélection des mutants EMS

Les mutants entourés sont ceux selectionnés pour ce projet. Le mutant A représente un mutant ayant une forte variabilité et le mutant B illuste un mutant ayant une faible variabilité comparée à celle du sauvage.

Chaque mutant EMS sélectionné a été séquencé et croisé avec la souche sauvage pour obtenir des cellules diploïdes hétérozygotes pour l'ensemble des mutations. Les levures haploïdes résultant de la méiose de ces diploïdes sont triées en 3 groupes (en anglais "bulk"), low, middle ou high, par FACS (Fluorescence Activated Cell Sorting) (voir Figure 4). Le FACS permet de trier les cellules selon leur fluorescence :

pour le low bulk, on a trié les ségrégants avec la plus faible fluorescence, pour le middle bulk les cellules avec une fluorescence proche de la moyenne et pour le high bulk les ségrégants avec la plus haute fluorescence.

Pour augmenter les chances de détecter la ou les mutations associées à la variabilité de fluorescence, trois étapes successives de sélection au FACS suivie de croissance des cellules ont été effectuées pour chaque bulk. Par exemple, les cellules du low bulk sont mises en culture (on passe de 150 000 cellules par bulk à 50 millions) puis triées au FACS et on conserve uniquement les cellules avec la plus faible fluorescence. Ces cellules sont mises en culture et encore une fois triées par FACS et on conserve les plus faiblement fluorescentes.

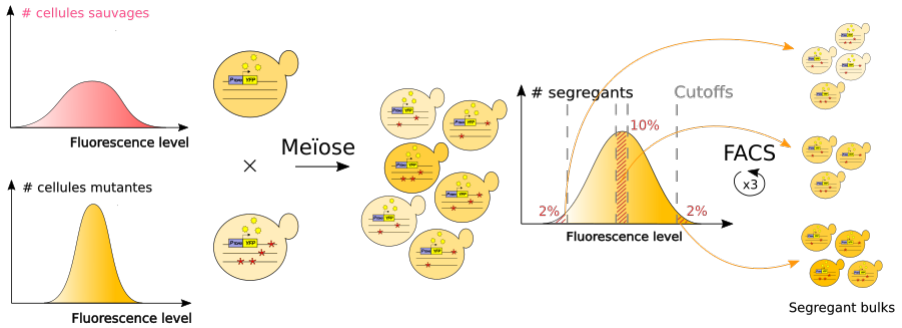


FIGURE 4 – Croisement entre une cellule sauvage et une cellule d'un mutant EMS puis triage des cellules par FACS (Fluorescence activated cell sorting)
Les cellules haploïdes issues de la méiose des diploïdes sont appelées "segregants". Ceux-ci sont triés par FACS en 3 bulks (low, middle, high) selon leur fluorescence.

Une méthode pour détecter des SNP consiste à comparer les séquences des cellules issues de la méiose des croisements aux séquences des cellules parents [2]. Dans ce projet, nous utilisons la méthode du BSA-Seq. Après avoir triées les cellules en plusieurs bulks, l'ADN génomique de tout le bulk est extrait et séquencé. Les mutations associées à la variabilité d'expression sont alors identifiées car l'allèle mutant n'apparaît pas à la même fréquence dans les différents bulks.

La Figure 5 explique comment à partir du séquençage des différents bulks et la fréquence d'allèle mutant, on détermine les mutations associées à la variabilité du phénotype. Cette figure illustre un cas où le mutant analysé par BSA-Seq présente deux mutations nommées a et b. Après le triage des segregants en trois bulks, chaque bulk est séquencé et on obtient la fréquence d'allèle pour chaque mutation et pour chaque bulk.

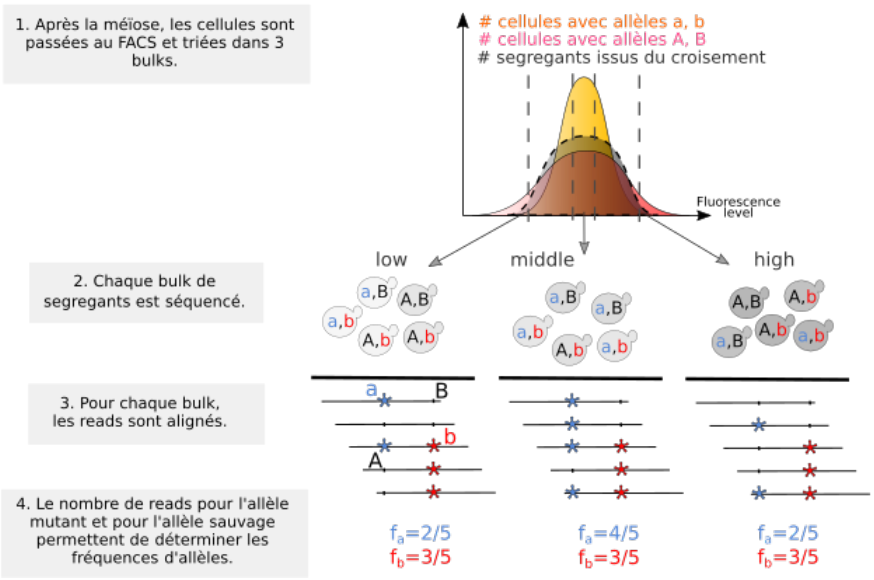


FIGURE 5 – Principe du BSA-Seq (Bulk Segregants Analysis and Sequencing) : Triage en bulk, séquençage et fréquence d'allèles

L'allèle mutant b est présent à la même fréquence dans tous les bulks, il n'a donc pas d'effet sur la variabilité. Les fréquences de l'allèle mutant a sont différentes dans les bulks low et high de celle du bulk middle. Le likelihood-test réalisé sur les comptes de reads sauvages/mutants par bulk permet de déterminer si cette différence est significative.

Dans cet exemple, la variabilité d'expression du mutant analysé par BSA-Seq est plus faible que celle du sauvage. Ceci implique que dans le cas où une mutation est associée à la variabilité d'expression du phénotype, la fréquence d'allèle mutant sera supérieure dans le bulk central que dans les bulks extrêmes. La fréquence de la mutation b est identique dans les 3 bulks, elle n'est donc pas associée à la variabilité d'expression du phénotype. La mutation a a une fréquence d'allèle mutant dans le bulk central supérieure aux fréquences dans les bulks extrêmes. On peut donc conclure que la mutation a est associée à la variabilité d'expression du phénotype.

Dans notre projet, chaque mutant analysé par BSA-Seq a une trentaine de mutations dans son génome. A partir de la fréquence d'allèle mutant de chaque mutation et déterminée dans chaque bulk, on pourra déduire les mutations associées à la variabilité.

En résumé, dans le cas où le mutant EMS a une variabilité d'expression plus faible que le sauvage, on s'attend à ce que les mutations impliquées dans la différence de fluorescence se situent dans le bulk central. La fréquence d'allèle mutant de la mutation associée à la variabilité sera donc supérieure dans le bulk central comparée à celles des bulks extrêmes. Au contraire, si le mutant EMS a une variabilité d'expression plus forte que le sauvage,

les mutations impliquées dans la différence de fluorescence devraient être dans les bulks extrêmes. La fréquence de l'allèle mutant de la mutation associée à la variabilité sera donc inférieure dans le bulk central comparée à celles des bulks extrêmes.

2.2 Traitement des séquences : Nextflow pipeline

Les données de BSA-Seq proviennent de l'Université du Michigan tandis que les données de séquençage des 6 mutants EMS proviennent du LBMC. Lors de mon stage, j'ai traité ces séquences informatiquement afin d'identifier toutes les mutations des mutants EMS et à partir des données de BSA-Seq, trouver celles qui sont associées à la variabilité d'expression.

En bioinformatique, on utilise plusieurs outils comme cutadapt, bowtie, freebayes *etc.* permettant chacun de réaliser une tâche spécifique ; par exemple supprimer les adaptateurs ou les séquences de mauvaise qualité, mapper contre le génome de référence ou filtrer les mutations. Ces outils fonctionnent indépendamment les uns des autres et nécessitent en entrée des fichiers avec un format spécifique, néanmoins ils s'appliquent à la chaîne, c'est-à-dire qu'après avoir enlevé les adaptateurs par exemple, on va utiliser les fichiers obtenus en sortie pour l'étape suivante "trimm per quality" (filtre les séquences selon leur qualité). Pour simplifier l'utilisation de ces outils, on utilise une pipeline qui permet de donner en entrée à l'outil suivant, les données de sortie de l'outil précédent.

Le logiciel Nextflow permet d'écrire des pipelines et garantit la reproductibilité des données grâce à son utilisation des conteneurs Docker ou Singularity. Un conteneur contient le code et toutes ses dépendances ce qui permet à l'application d'être exécutée de manière fiable d'un environnement informatique à un autre. On utilise donc deux fichiers pour Nextflow, un pour le code et un pour paramétrer les conteneurs. L'ENS de Lyon dispose d'un Pôle Scientifique de Modélisation Numérique (PSMN) sur lesquel on peut utiliser les ressources en calcul haute performance pour lancer notre pipeline et récupérer les fichiers en sortie.

La Figure 6 est une réprésentation de la pipeline Nextflow écrite pour les données de séquençage de ce projet. Cette pipeline est adaptable pour des données similaires (séquences de mutants ou BSA-seq) et est disponible sur le gitlab de l'ENS de Lyon.

Les séquences des mutants EMS ont été obtenues par séquençage MiSeq et les reads pairés font 2*250 bp. Les séquences des bulks résultent du séquençage Illumina et les reads pairés font 2*150 bp. La pipeline prend en compte ces différences, notamment lors de la

suppression des séquences d'adaptateurs et lors de l'identification des SNP.

A chaque étape de la pipeline, les outils bioinformatiques disposent d'options (minimum de qualité *etc.*) que nous avons ajusté selon les résultats obtenus dans les MultiQC. Ces derniers présentent la qualité des résultats à chaque étape de manière synthétique.

En entrée, on doit spécifier un fichier fasta contenant le génome de référence et un fichier csv contenant l'identifiant du mutant, la population de segregants (low, middle ou high dans le cas de données BSA-Seq) et le chemin des paires de fichiers fastq.

Une fois que toutes les étapes ont été réalisées avec succès, on obtient des fichiers textes et des fichiers avec un format spécifique : le Variant Call Format (VCF). Dans chacun des VCF, on dispose de la liste de toutes les mutations détectées informatiquement.

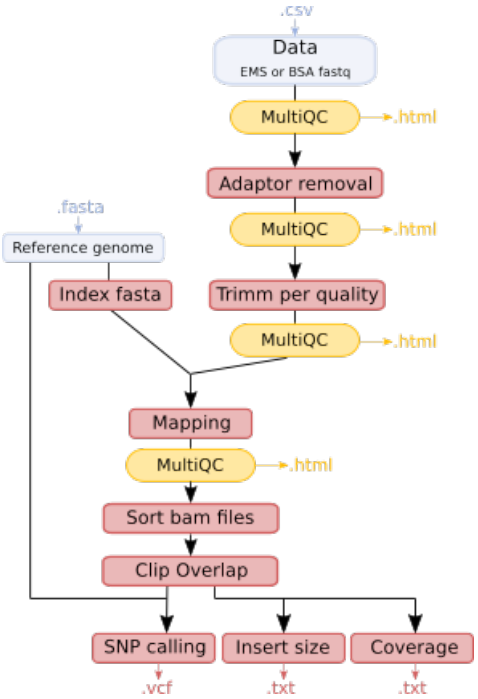


FIGURE 6 – Schéma de la pipeline d'analyse

La pipeline prend en entrée un fichier csv et un fichier fasta. A chaque étape en rouge, un outil bioinformatique traite les données. Les MultiQC sont des rapports synthétiques de la qualité des résultats obtenus après traitement par l'outil.

L'analyse est différente pour les mutants EMS et pour les mutants analysés par BSA-Seq lors de l'étape d'identification des propositions (en anglais SNP calling). On parle de propositions car les propositions ne sont pas toutes des mutations. En effet, certaines propositions sont des artefacts dûs au séquençage. Dans le cas où l'on traite les mutants analysés par BSA-Seq, on aimerait connaître le nombre de reads contenant l'allèle mutant (RO) et le nombre de reads contenant l'allèle sauvage (AO), pour chaque bulk. L'outil freebayes permet de gérer ces spécificités et renvoie (pour chaque mutant analysé par BSA-Seq) un fichier vcf contenant des informations par bulk (voir Figure 10 en annexe).

2.3 Filtrer les mutations chez les mutants EMS

On ne s'intéresse qu'à une partie de l'ensemble des propositions détectées dans les mutants EMS par l'outil freebayes. On souhaite conserver parmi l'ensemble des proposi-

tions celles répondant à nos critères (SNP, score de qualité supérieur à 200, profondeur de séquençage supérieure à 5 *etc.*). Le script *filter_ vcf.py* renvoie pour chaque fichier vcf en entrée, un fichier vcf en sortie contenant les mutations vérifiant l'ensemble de nos filtres.

Le script *filter_ occurances.py* compte le nombre de fois qu'une mutation apparaît dans l'ensemble des fichiers vcf. Certaines mutations apparaissent dans tous les fichiers vcf et ne sont donc pas impliquées dans la différence de fluorescence. Elles peuvent provenir d'une différence entre le génome de la souche sauvage utilisée comme référence et le génome de la souche sauvage à l'origine des mutants EMS, ce qui expliquerait pourquoi ces mutations se retrouvent dans l'ensemble des mutants EMS. Dans les 9 fichiers vcf, on trouve de nombreuses mutations ayant une occurrence de 9 mais quelques mutations ont une occurence de 7 ou 8. On considère que ces mutations sont présentes pour la même explication mais qu'il y a quelques erreurs de séquençage. Certaines mutations ont une occurrence de 4 ou 5 et proviennent problablement de régions difficiles à séquencer et donc sont difficilement détectables. Le script $filte\_occurances.py$ renvoie pour chaque fichier vcf en entrée, un fichier vcf en sortie contenant uniquement les mutations ayant une occurrence égale à 1. Ce script imprime également le nombre de SNP par fichier de sortie et la proportion de C mutée en T et G en A.

Le script bash *treatment_ EMS_ vcf.sh* exécute les deux scripts python et renvoie un fichier texte "results.txt" contenant les impressions.

2.4 Filtrer les mutations chez les segregants

Les segregants étant issus de la méiose du croisement entre un mutant EMS et la souche sauvage, on s'attend à retrouver dans les mutants analysés par BSA-Seq, les mutations présentes dans le mutant EMS. Le but de l'analyse par BSA-Seq est d'identifier parmi les mutations présentes dans les mutants EMS, lesquelles sont associées à une différence de variabilité. Parmi l'ensemble des propositions de freebayes, on garde uniquement les mutations identifiées dans le mutant EMS.

Le script *compare_ EMS_ BSA.py* permet d'identifier les mutations communes (mutations apparaissant dans le mutant EMS et dans le mutant analysé par BSA-Seq (issu du croisement de ce même mutant et de la souche sauvage)) et note également les mutations manquantes (mutations présentes dans le mutant EMS mais absentes dans le mutant analysé par BSA-Seq correspondant). Le script renvoie dans un fichier texte les mutations communes, le nombre de reads contenant l'allèle sauvage et le nombre de reads contenant

l'allèle mutant pour chacun des bulks (voir Figure 11 en annexe).

2.5 Identification et tests statistiques

A cette étape du projet, on dispose pour les 3 bulks, du nombre de reads contenant l'allèle sauvage (RO) et du nombre de reads contenant l'allèle mutant (AO) pour chaque mutation détectée dans les mutants analysées par BSA-Seq. Ces données sont présentées dans une table de contingence (voir tableau 1).

TABLE 1 – Table de contingence pour le likelihood test à 3 bulks (gauche) et à 2 bulks (droite)

	Low	Middle	High
RO	.	.	.
AO	.	.	.

	Low+High	Middle
RO	.	.
AO	.	.

On effectue un premier test du rapport de vraisemblance ('likelihood ratio test'). Dans notre cas, ce test calcule 1) la probabilité que les fréquences observées dans chacun des bulks (low, middle, high) suivent une même loi de probabilité et 2) la probabilité que les comptes (AO et RO) et les bulks sont indépendants. G est calculé de la façon suivante : $G = 2\sum ln(f_{observes}/f_{attendues})$ avec $f_{observes}$ les fréquences observées et $f_{attendues}$ les fréquences attendues. Pour de grands échantillons, G suit une loi du χ^2 à un degré de liberté. A partir des tables, on déduit la pvalue associée à la valeur de G [4] Voir détails ici. Dans notre cas, on a utilisé les fonctions R *likelihood.ratio* et *chisq.test* pour obtenir les pvalues.

On effectue un autre test du rapport de vraisemblance testant cette fois : 1) la probabilité que la fréquence observée dans les bulks extrêmes (low + high) et la fréquence observée dans le bulk middle suivent une même loi de probabilité et 2) la probabilité que les comptes (AO et RO) et les bulks sont indépendants.

En statistiques, lorsque l'on compare de multiples échantillons, il faut penser à appliquer une méthode d'ajustement permettant d'ajuster le seuil de significativité (appelé le risque α). En effet, plus on effectue de tests, plus ce risque augmente (on parle d'inflation du risque α). Le risque de conclure à tort à la réalité est donc très fortement augmenté. On a choisi d'appliquer la correction de Benjamini & Hochberg (BH) aux pvalues calculées par le likelihood ratio test. On fixe le seuil de significativité à 0,01. Les mutations ayant une pvalue inférieure à ce seuil sont dites significatives.

Pour une meilleure représentation des résultats, on représente le pscore pour chaque mutation. Pour le test du rapport de vraisemblance sur les 3 bulks, le pscore est égal à $-\log_{10}(pvalue_{3bulks})$. Pour le test du rapport de vraisemblance sur les 2 bulks (low+high et middle), le pscore est égal à $-\log_{10}(pvalue_{2bulks}) * (\pm 1)$. Si la fréquence du nombre de reads contenant l'allèle mutants dans le bulk low et high est supérieure ou égale à la fréquence du nombre de reads contenant l'allèle mutant dans le bulk middle alors on multiplie par 1, sinon par -1. En effet, dans le cas où le mutant analysé par BSA-Seq a une variabilité d'expression plus forte que le sauvage, la fréquence d'allèle mutant de la mutation associée à la différence de variabilité doit être supérieure dans le bulk low+high. Dans ce cas, on considère que les mutations avec un pscore positif supérieur à 2 sont significativement associées à une augmentation de variabilité d'expression. Dans le cas, où le mutant analysé par BSA-Seq a une variabilité plus faible que le sauvage, la fréquence d'allèle mutant de la mutation associée à la différence de variabilité doit être dans le bulk middle. On considère que les mutations avec un pscore négatif inférieur à -2 sont significativement associées à une diminution de variabilité d'expression.

3 Résultats et Discussion

3.1 Qualité des séquences et alignement

Les fichiers MultiQC permettent d'obtenir des statistiques sur la qualité des séquences et l'alignement des différents fichiers fastq traités. Cette partie présente quelques résultats issus des MultiQC.

Données de séquençage des mutants EMS

On dispose de 9 fichiers fastq correspond chacun à un mutant EMS. On s'intéressera uniquement à 6 d'entre eux pour ce projet mais les statistiques suivantes sont réalisées sur les 9 mutants EMS séquencés.

Chaque paire de données contient en moyenne 3095106 reads (minimum : 2628224, maximum : 3918041). Le Phred Score (qualité moyenne par base pour chaque read) était supérieur à 30 avant et après l'étape cutadapt, et est supérieur à 35 après l'étape trimm.

Bowtie 2 a été utilisé pour l'alignement des reads avec le génome de référence. Les reads sont alignés en moyenne à 97,82%. Pour chaque paire, on observe en moyenne 1384872 (minimum : 1177180, maximum : 1819884) reads alignés de façon unique au génome de

référence.

On conclut que la qualité des séquences des mutants EMS est satisfaisante et le pourcentage d'alignement des reads est bon.

Données de séquençage des mutants analysés par BSA-Seq

On dispose de 32 fichiers fastq (18 paires). Le MultiQC avant l'étape cutadapt confirme que 2 paires de fichiers sont erronées (voir Figure 12 en annexe). En effet, une paire contient 18 258 reads tandis qu'une autre en contient 31 806 994. En moyenne, les paires de fichiers contiennent 10M de reads. Les paires de fichiers erronées correspondent au même segregant. On exclut ce segregant (2P20G07) de l'analyse car les fichiers contiennent trop peu ou trop de reads. Ceci s'explique probablement par une erreur lors du multiplexage des échantillons.

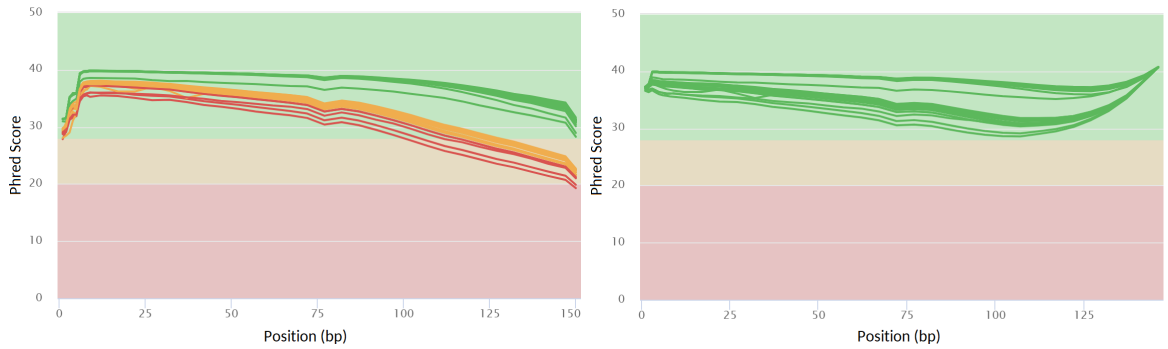


FIGURE 7 – Scores moyens de qualité avant et après trimm

Scores moyens de qualité par bp pour chaque fichier fastq avant et après trimm

Bowtie 2 a été utilisé pour l'alignement des reads avec le génome de référence. Les reads sont correctement alignés en moyenne à 95.85% (minimum : 94,7%, maximum : 97%). Pour chaque paire, on observe en moyenne 2500000 reads alignés de façon unique au génome de référence.

Les scores moyens de qualité avant le trimm sont compris entre 19 et 40 alors qu'après le trimm, ils sont entre 30 et 40 (voir Figure 7). Cette étape de la pipeline permet donc comme attendu de supprimer les séquences de mauvaise qualité.

3.2 Mutations identifiées dans les mutants EMS

TABLE 2 – Nombre de SNP identifiées pour chaque mutant EMS et proportion de SNP C en T et G en A

Mutants EMS	2P05A05	2P09D12	2P15E12	2P16H04	2P17F03	2P20G07
Nombre de SNP	17	16	26	47	23	30
C :T et G :A (en %)	100	93,7	100	100	100	96,66

Le tableau 2 présente le nombre de mutations identifiées par mutant EMS et parmi ces mutations, la proportion de bases C mutées en T et G en A. Les mutants étudiés lors de ce projet proviennent de l'étude de 1922 mutants ([5]) et Metzger *et al.* avaient estimé que le nombre de mutations par mutant serait de 32. Ceci est tout à fait cohérent avec nos observations puisqu'en moyenne un mutant EMS présente 26,5 mutations. On peut également conclure que la quantité d'EMS a été correctement dosée. De plus, on observe que la proportion moyenne de bases C mutées en T et G en A est de 98,39%. L'EMS étant un mutagène chimique induisant à 96% des mutations des bases C en T et G en A, on conclut que les mutations identifiées sont dues à l'EMS.

3.3 Mutations identifiées dans les mutants analysés par BSA-Seq

TABLE 3 – Nombre de SNP identifiées dans chaque mutant analysé par BSA-Seq

Mutant analysé par BSA-Seq issu de la souche sauvage et du mutant :	2P05A05	2P16H04	2P09D12	2P15E12	2P17F03
Nombre de mutations communes	17	47	16	26	21
Nombre de mutations manquantes	0	0	0	0	2

Les mutations identifiées dans un mutant EMS doivent apparaître dans le génome du mutant analysé par BSA-Seq puisqu'il est issu du croisement entre ce même mutant EMS et la souche sauvage. Le tableau 3 présente les mutations communes et manquantes identifiées dans les mutants analysés par BSA-Seq. Pour 4 mutants analysés par BSA-Seq sur 5, toutes les mutations présentes dans le mutant EMS ont bien été identifiées dans le mutant analysé par BSA-Seq correspondant. Pour le mutant analysé par BSA-Seq issu du mutant 2P17F03, 2 mutations présentes dans le mutant EMS n'apparaissent pas dans le génôme du mutant analysé par BSA-Seq.

TABLE 4 – Fréquence de l'allèle mutant pour les 2 mutations du mutant EMS 2P17F03 n'apparaissant pas dans le mutant analysé par BSA-Seq correspondant

chr09 380951 C T			
	RO	AO	freq
Low	15	23	≈ 0,61
Middle	14	18	≈ 0,56
High	21	32	≈ 0,60

chr14 278253 G A			
	RO	AO	freq
Low	104	3	≈ 0,03
Middle	103	2	≈ 0,02
High	150	1	≈ 0,01

Le tableau 4 présente les fréquences de l'allèle mutant pour ces 2 mutations manquantes. Ces fréquences ont été obtenues à l'aide de l'outil Integrative Genomics Viewer (IGV) qui permet de visualier l'alignement des reads contre le génome de référence. On compte alors le nombre de reads contenant l'allèle sauvage (RO) et le nombre de reads contenant l'allèle mutant (AO) pour les 2 mutations souhaitées. La fréquence de l'allèle mutant est calculée comme suit : AO/(RO+AO).

La mutation sur le chromosome 9 se situe dans une région non codante avant le gène VLD1. Les fréquences d'allèles calculées pour chaque bulk de la mutation sur le chromosome 9 ne semblent pas suggérer une différence statistique significative entre les fréquences d'allèles des bulks extrêmes et milieu. On peut considérer que cette mutation n'est pas associée à la variabilité d'expression.

Il est important d'observer que le nombre de reads à cette position est extrêmement faible (seulement une trentaine de reads pour une couverture estimée à 100). Pour comprendre cette observation, nous avons visualisé la couverture des mutants EMS et la couverture des mutants analysés par BSA-Seq pour chaque bulk. On constate que les mutants analysés par BSA-Seq ont une couverture plus variable que les mutants EMS. Les reads n'étant pas de la même taille, 2*250 bp pour les mutants EMS et 2*150 bp pour les segregants, on peut supposer que les reads de taille plus grande sont plus précisement alignés contre le génome. Les reads de taille plus petite peuvent s'aligner à plus d'endroits et expliquent les fluctuations observées dans la couverture des mutants analysés par BSA-Seq.

La mutation sur le chromosome 14 se situe dans le gène CHS1. Les fréquences d'allèles dans les 3 bulks sont très faibles. Il est possible que cette mutation ait été contre-selectionnée. Par exemple, si la mutation touche un gène important pour la méiose ou la prolifération, on peut s'attendre à ce qu'elle soit délétère et donc contre-sélectionnée. La mutation sur le chromosome 14 à la position 278253 induisant une base A a très probable-

ment été contre-selectionnée. Ceci expliquerait pourquoi on observe si peu fréquemment l'allèle mutant.

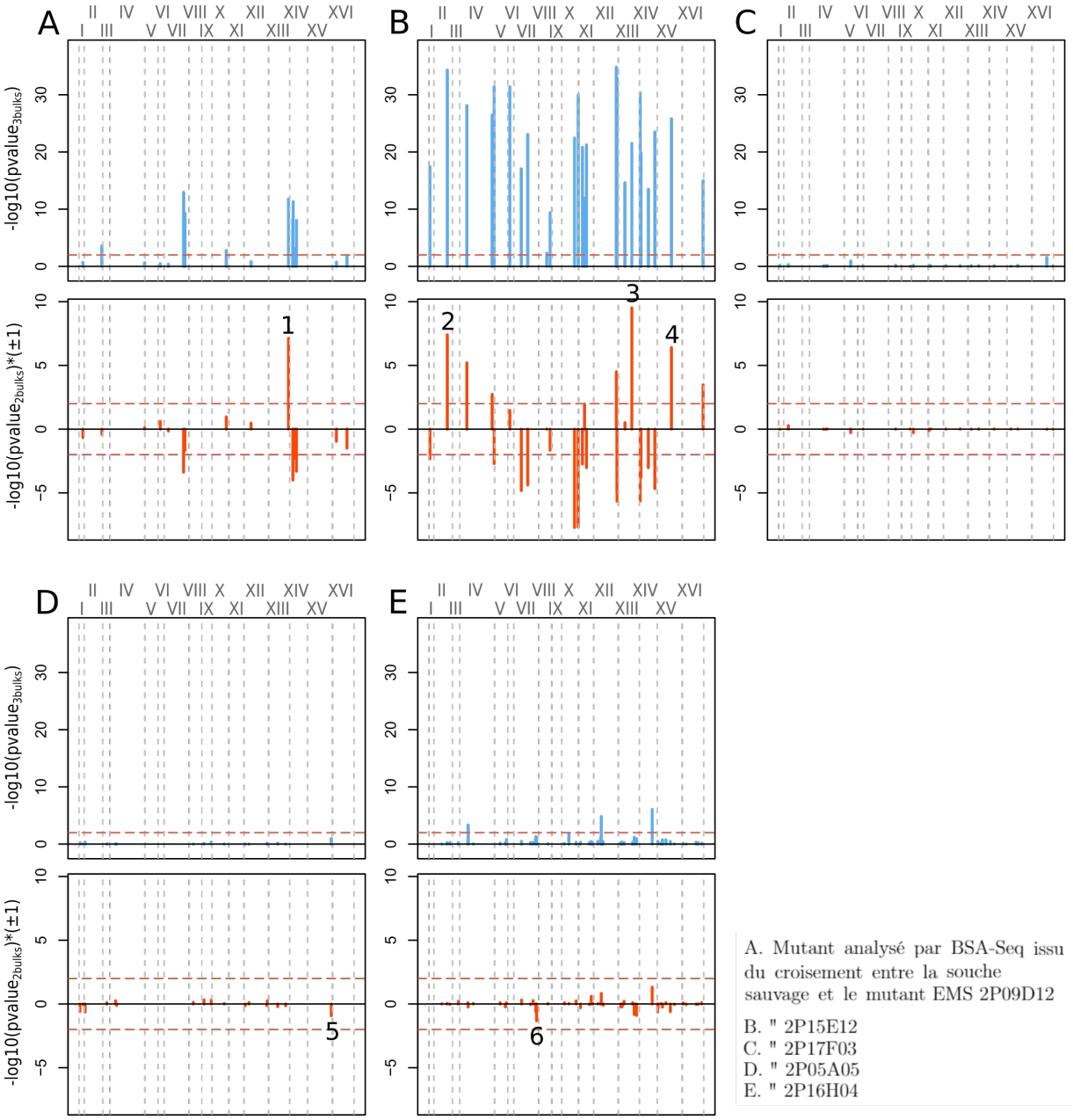
3.4 Mutations contre-sélectionnées ou sélectionnées

TABLE 5 – Mutations contre-selectionnées ou selectionnées

Chromosome	Position	Réf.	Alt.	Fréquence de l'allèle mutant	Présent dans le gène
chr03	175878	C	T	0.938	SYP1
chr08	350835	G	A	0.06	MSH1
chr08	495174	G	A	0.23	RIX1
chr14	215425	G	A	0.15	CSL4

Pour tester cette hypothèse de contre-sélection ou sélection, nous avons calculé pour chaque mutation (dans l'ensemble des mutants analysés par BSA-Seq), la fréquence de l'allèle mutant. En l'absence de sélection, pour un organisme haploïde, 50% des reads contiennent l'allèle mutant et 50% des reads contiennent l'allèle sauvage. Le tableau 5 présente les mutations ayant une fréquence très différente de 50%. Toutes ces mutations sont présentes dans des gènes. On observe donc que certaines mutations sont contre-selectionnées (mutations sur le chromosome 8 et 14 du tableau 5) ou selectionnées (mutation sur le chromosome 3 du tableau 5).

3.5 Tests statistiques



A. Mutant analysé par BSA-Seq issu du croisement entre la souche sauvage et le mutant EMS 2P09D12

B. " 2P15E12
C. " 2P17F03
D. " 2P05A05
E. " 2P16H04

FIGURE 8 – Pscores pour chaque mutation identifiée dans les mutants analysés par BSA-Seq (A, B et C ont une variabilité plus forte que le sauvage, et D et E ont une variabilité plus faible)

En bleu sont représentés les pscores obtenus pour le likelihood ratio test des 3 bulks (low, middle, high) et en orange les pscores obtenus pour le likelihood ratio test des 2 bulks (low+high, middle). La ligne rouge horizontale correspond à une pvalue de 0.01.

La Figure 8 présente les pscores obtenus pour chaque mutation identifiée dans les mutants analysés par BSA-Seq. Si on observe tout d'abord les 3 mutants analysés par

BSA-Seq ayant une variabilité d'expression plus forte que le sauvage, on constate que le nombre de mutations significatives (c'est-à-dire pscore en bleu au dessus de 2 et en orange au dessus de 2) est très variable d'un mutant à l'autre. Le mutant 2P05A05 analysé par BSA-Seq a une même mutation significative à la fois pour le test 3bulks et le test 2bulks. Le mutant 2P15E12 analysé par BSA-Seq a 7 mutations significatives pour les deux tests et enfin le mutant 2P17F03 analysé par BSA-Seq n'a aucune mutation significative pour les deux tests. Cette observation est assez surprenante car dans l'expérience de détection des mutations associées à la différence de moyenne [1], une mutation très significative été détectée par mutant analysé par BSA-Seq.

L'unique mutation significative dans le mutant 2P09D12 analysé par BSA-Seq est donc une mutation très problablement associée à la différence de variabilité d'expression (on parlera de mutation candidate). Lors de l'analyse du mutant 2P15E12 par BSA-Seq, il se peut qu'il y ait eu de fortes réductions de taille de population. En effet, si la population de segregants diminue (lors des tris au FACS par exemple), les variations aléatoires de fréquence d'allèles augmentent par dérive génétique et les fréquences d'allèles ne sont plus représentatives de la sélection appliquée par FACS car il n'y a pas assez de segregants. On observe donc plus de mutations significatives car il y a moins de segregants dans l'échantillon et que les variations de fréquences sont donc plus significatives. Parmi les mutations significatives pour le mutant 2P15E12 analysé par BSA-Seq, on considère les trois mutations avec le plus haut pscore comme des mutations candidates. Le mutant 2P17F03 analysé par BSA-Seq n'a aucune mutation significative pour les deux tests. Il s'agit du mutant analysé par BSA-Seq où deux mutations étaient manquantes. La mutation associée à la différence de variabilité d'expression peut être parmi ces deux mutations manquantes. Nous avons vu que l'une d'entre elle (chr14 278253 G A) avait probablement été contre-selectionnée mais nous ne pouvons pas exclure qu'elle soit associée à la différence de variabilité, elle est donc aussi une mutation candidate.

Pour les 2 mutants analysés par BSA-Seq ayant une variabilité d'expression plus faible que le sauvage, on constate qu'il n'y a pas de mutation significative pour les deux tests (pscore en bleu au dessus de 2 et en orange en dessous de -2). Plusieurs hypothèses peuvent expliquer l'abscence de mutation significative. Tout d'abord, nous avons vu que les mutations peuvent être contre-sélectionnées. Ceci explique que le nombre de reads contenant l'allèle mutant soit très faible et que les fréquences d'allèle mutant soient proches de 0. Il n'est alors pas possible de détecter la mutation. Une autre explication peut être que les conditions de culture étaient différentes lors du tri par FACS (en tube) et lors de

l'analyse des mutants EMS (en plaque 96 puits). Les conditions expérimentales n'étant pas les mêmes, il se peut donc que l'environnement ait joué un rôle sur l'expression. La différence de variabilité d'expression entre la souche sauvage et la souche mutante n'est peut être pas due à une mutation mais à des modifications épigénétiques. On s'intéresse tout de même aux mutations (5 et 6) avec les pscores les plus négatifs et on les ajoute à la liste des mutations candidates car il est possible qu'elles aient un effet sur la variabilité mais que celui-ci soit trop faible pour être détecté par BSA-Seq.

TABLE 6 – Liste des mutations candidates et informations complémentaires
Remarque : la mutation 7 est une mutation manquante dans le mutant analysé par BSA-Seq (issu du croisement de la souche sauvage et du mutant EMS 2P17F03) mais tout de même retenue comme mutation candidate.

	Chrom	Pos	Réf	Alt	Présent dans le gène	Modification de l'acide aminé
1	chr13	871392	C	T	YME2	D → N
2	chr02	582914	G	A	SWD3	E → E
3	chr13	590301	G	A	PAH1	A → A
4	chr15	618727	G	A	/	/
5	chr15	1039495	C	T	/	/
6	chr07	995063	C	T	/	/
7	chr14	278253	G	A	CHS1	W → STOP

Le tableau 6 résume pour chaque mutation candidate, sa présence dans un gène ou non, et la modification de l'acide aminé qu'elle entraîne ou non. Les mutations candidates 1,2,3 et 7 sont toutes présentes dans un gène tandis que les mutations candidates 4,5 et 6 sont situées dans des régions intergéniques. Les mutations candidates 2 et 3 sont des mutations synonymes c'est-à-dire qu'elle ne modifie pas l'acide aminé. La mutation candidate 7 est une mutation non-sens, c'est-à-dire que le codon codant pour l'acide aminé est remplacé par un codon stop. Enfin, la mutation candidate 1 modifie l'acide aspartique (D) en un asparagine (N).

3.6 Fréquence d'allèles pour 6 mutations candidates

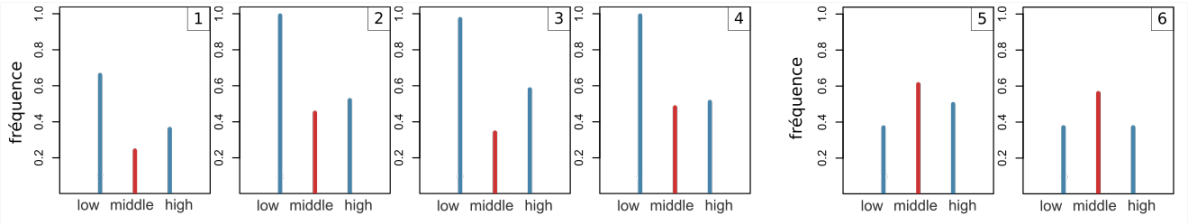


FIGURE 9 – Fréquence d'allèles mutants dans chaque bulk pour différentes mutations candidates

Les mutations 1 à 4 sont issues de mutants analysés par BSA-Seq avec une variabilité plus forte que le sauvage et 5 et 6 de mutants analysés par BSA-Seq avec une variabilité plus faible que le sauvage.

La Figure 9 représente les fréquences de l'allèle mutant dans chaque bulk pour chaque mutation candidate. Pour les mutations candidates 1 à 4, on observe comme attendu que les fréquences d'allèles des bulks extrêmes (en bleu) sont supérieures à la fréquence du bulk central (en rouge). Les mutations 5 et 6 ont une fréquence d'allèle du bulk central supérieure à la fréquence d'allèle du bulk low et du bulk high. Ces observations sont cohérentes avec la théorie puisque dans le cas où le mutant EMS a une variabilité d'expression plus faible que le sauvage, la probabilité que la mutation associée à la variabilité d'expression de fluorescence se trouve dans le bulk central est supérieure à la probabilité qu'elle soit dans les bulks extrêmes. La fréquence d'allèle dans le bulk central doit donc être supérieure à celles des autres bulks.

4 Conclusion

Ce projet a tout d'abord permis de développer une pipeline d'analyse de données de séquences d'ADN génomique et de données de BSA-Seq. Ainsi, l'ensemble des mutations présentes dans les 6 souches mutantes ont été identifiées. De plus, l'ensemble des mutations présentes dans les données de BSA-Seq ont également été identifiées et comparées avec celles des souches mutantes. Pour chaque mutation commune, nous avons déterminé la fréquence d'allèle dans chaque bulk et enfin utilisé une approche statistique pour identifier les mutations associées à la différence de variabilité d'expression entre la souche sauvage et la souche mutée. Les mutations résultant de cette analyse, les mutations candidates, seront chacune vérifiée par introgression, c'est-à-dire réinjecter chacune dans le génome

sauvage. Cette méthode utilisée également lors de l'identification de mutations associée à la différence d'expression en moyenne [1] permet de vérifier que la mutation candidate est associée au changement de variabilité d'expression observée dans le mutant EMS.

Si les résultats de l'introgression permettent de conclure que le protocole ainsi que le traitement des données ont permis de détecter avec succès des mutations associées à la variabilité d'expression, une possibilité serait de refaire l'expérience avec un plus grand nombre de mutants EMS et d'établir si les mutations détectées correspondent à des gènes en particulier.

Identifier les mutations associées aux changements de variabilité d'expression permet d'accroître notre compréhension de la diversité des mécanismes cellulaires impliqués dans la variabilité. Ceci permettrait également de mieux comprendre de nombreux processus fondamentaux liés à la variabilité tels que le développement, la persistance bactérienne ou la cancérogénèse.

Références

[1] Fabien Duveau, Brian PH Metzger, Jonathan D Gruber, Katya Mack, Natasha Sood, Tiffany E Brooks, and Patricia J Wittkopp. Mapping small effect mutations in saccharomyces cerevisiae : impacts of experimental design and mutational properties. *G3 : Genes, Genomes, Genetics*, 4(7) :1205–1216, 2014.

[2] Steffen Fehrmann, Hélène Bottin-Duplus, Andri Leonidou, Esther Mollereau, Audrey Barthelaix, Wu Wei, Lars M Steinmetz, and Gaël Yvert. Natural sequence variants of yeast environmental sensors confer cell-to-cell expression variability. *Molecular Systems Biology*, 9(1) :695, 2013.

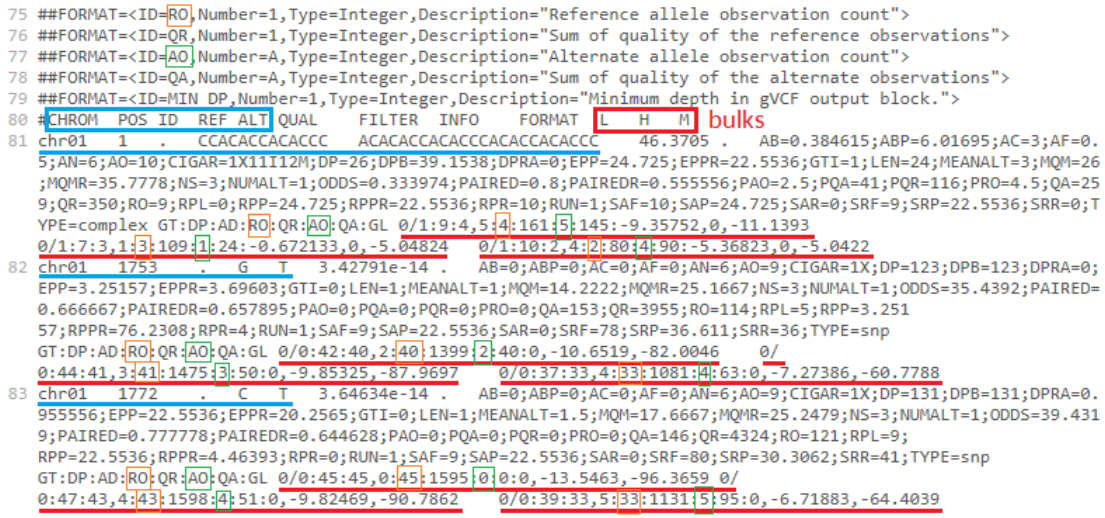
[3] Gil Hornung, Raz Bar-Ziv, Dalia Rosin, Nobuhiko Tokuriki, Dan S Tawfik, Moshe Oren, and Naama Barkai. Noise–mean relationship in mutated promoters. *Genome research*, 22(12) :2409–2417, 2012.

[4] Influentialpoints.com. The g likelihood-ratio test [consulté le 12 septembre 2020].

[5] Brian PH Metzger, Fabien Duveau, David C Yuan, Stephen Tryban, Bing Yang, and Patricia J Wittkopp. Contrasting frequencies and effects of cis-and trans-regulatory mutations affecting gene expression. *Molecular biology and evolution*, 33(5) :1131–1146, 2016.

[6] Sydney M Shaffer, Margaret C Dunagin, Stefan R Torborg, Eduardo A Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A Brafford, Min Xiao, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658) :431–435, 2017.

Annexes




```
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  L   H   M
chr01   1   .   CCACACCACACCC   ACACACCACACCCACACCACACCC    46.3705 .   AB=0.384615;ABP=6.01695;AC=3;AF=0.
5;AN=6;AO=10;CIGAR=1X11I12M;DP=26;DPB=39.1538;DPRA=0;EPP=24.725;EPPR=22.5536;GTI=1;LEN=24;MEANALT=3;MQM=26
;MQMR=35.7778;NS=3;NUMALT=1;ODDS=0.333974;PAIRED=0.8;PAIREDR=0.555556;PAO=2.5;PQA=41;PQR=116;PRO=4.5;QA=25
9;QR=350;RO=9;RPL=0;RPP=24.725;RPPR=22.5536;RPR=10;RUN=1;SAF=10;SAP=24.725;SAR=0;SRF=9;SRP=22.5536;SRR=0;T
YPE=complex GT:DP:AD:RO:QR:AO:QA:GL 0/1:9:4,5:4:161:5:145:-9.35752,0,-11.1393
0/1:7:3,1:3:109:1:24:-0.672133,0,-5.04824    0/1:10:2,4:2:80:4:90:-5.36823,0,-5.0422
chr01   1753    .   G   T   3.42791e-14 .   AB=0;ABP=0;AC=0;AF=0;AN=6;AO=9;CIGAR=1X;DP=123;DPB=123;DPRA=0;
EPP=3.25157;EPPR=3.69603;GTI=0;LEN=1;MEANALT=1;MQM=14.2222;MQMR=25.1667;NS=3;NUMALT=1;ODDS=35.4392;PAIRED=
0.666667;PAIREDR=0.657895;PAO=0;PQA=0;PQR=0;PRO=0;QA=153;QR=3955;RO=114;RPL=5;RPP=3.251
57;RPPR=76.2308;RPR=4;RUN=1;SAF=9;SAP=22.5536;SAR=0;SRF=78;SRP=36.611;SRR=36;TYPE=snp
GT:DP:AD:RO:QR:AO:QA:GL 0/0:42:40,2:40:1399:2:40:0,-10.6519,-82.0046    0/
0:44:41,3:41:1475:3:50:0,-9.85325,-87.9697    0/0:37:33,4:33:1081:4:63:0,-7.27386,-60.7788
chr01   1772    .   C   T   3.64634e-14 .   AB=0;ABP=0;AC=0;AF=0;AN=6;AO=9;CIGAR=1X;DP=131;DPB=131;DPRA=0.
955556;EPP=22.5536;EPPR=20.2565;GTI=0;LEN=1;MEANALT=1.5;MQM=17.6667;MQMR=25.2479;NS=3;NUMALT=1;ODDS=39.431
9;PAIRED=0.777778;PAIREDR=0.644628;PAO=0;PQA=0;PQR=0;PRO=0;QA=146;QR=4324;RO=121;RPL=9;
RPP=22.5536;RPPR=4.46393;RPR=0;RUN=1;SAF=9;SAP=22.5536;SAR=0;SRF=80;SRP=30.3062;SRR=41;TYPE=snp
GT:DP:AD:RO:QR:AO:QA:GL 0/0:45:45,0:45:1595:0:0:0,-13.5463,-96.3659 0/
0:47:43,4:43:1598:4:51:0,-9.82469,-90.7862    0/0:39:33,5:33:1131:5:95:0,-6.71883,-64.4039
```

FIGURE 10 – Extrait d'un fichier vcf obtenu pour un mutant analysé par BSA-Seq avant comparaison avec les mutations du mutant EMS correspondant

Au début de chaque fichier vcf, on retrouve des informations (ligne de commande utilisée pour générer le vcf, légende etc.). Ensuite pour chaque ligne on a une mutation détectée (en bleu) et le nombre de reads contenant l'allèle sauvage (RO) (orange) et l'allèle mutant (AO) (vert) pour chaque bulk (en rouge)

CHROM	POS	REF	ALT	RO:L	AO:L	RO:M	AO:M	RO:H	AO:H
chr01	39905	G	A	70	59	61	77	49	38
chr01	169673	C	T	69	58	75	66	53	47
chr02	35745	C	T	38	33	36	51	34	22
chr03	175878	C	T	8	107	9	123	2	61
chr04	244028	G	A	74	70	82	61	51	50
chr04	290366	G	A	63	64	58	67	48	53
chr08	188596	G	A	56	60	74	67	42	42
chr09	94820	G	A	46	45	54	37	27	31
chr09	411056	G	A	80	69	73	54	38	52
chr10	540078	G	A	66	48	58	42	53	42
chr12	55033	C	T	81	71	74	77	53	46
chr12	213784	C	T	33	34	36	30	28	26
chr12	995410	C	T	64	55	55	64	38	47
chr12	1018632	C	T	71	65	63	46	26	31
chr13	395727	C	T	85	54	64	60	46	43
chr13	745307	C	T	68	43	63	46	42	23
chr15	1039495	C	T	63	37	30	47	38	38

FIGURE 11 – Liste des mutations identifiées chez le mutant analysé par BSA-Seq (issu du croisement entre la souche sauvage et le mutant 2P05A05) et communes aux mutations du mutant 2P05A05

Pour chaque mutation et pour chaque bulk (low : L, middle : M, high : H) on dispose du nombre de reads avec l'allèle sauvage (RO) et du nombre de reads avec l'allèle mutant (AO)

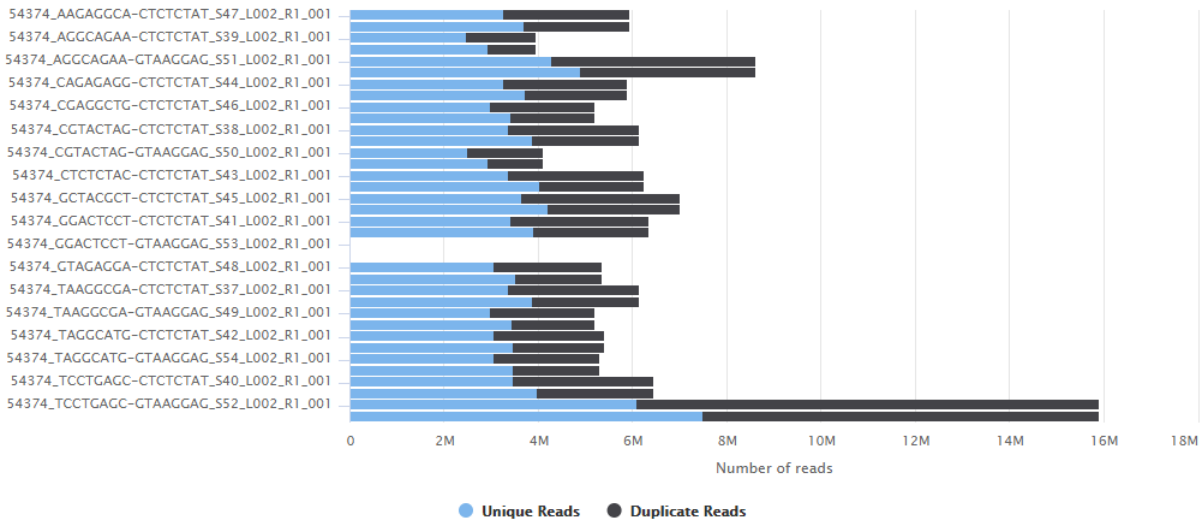


FIGURE 12 – FasQC Sequence count

Taux de couverture des segregants

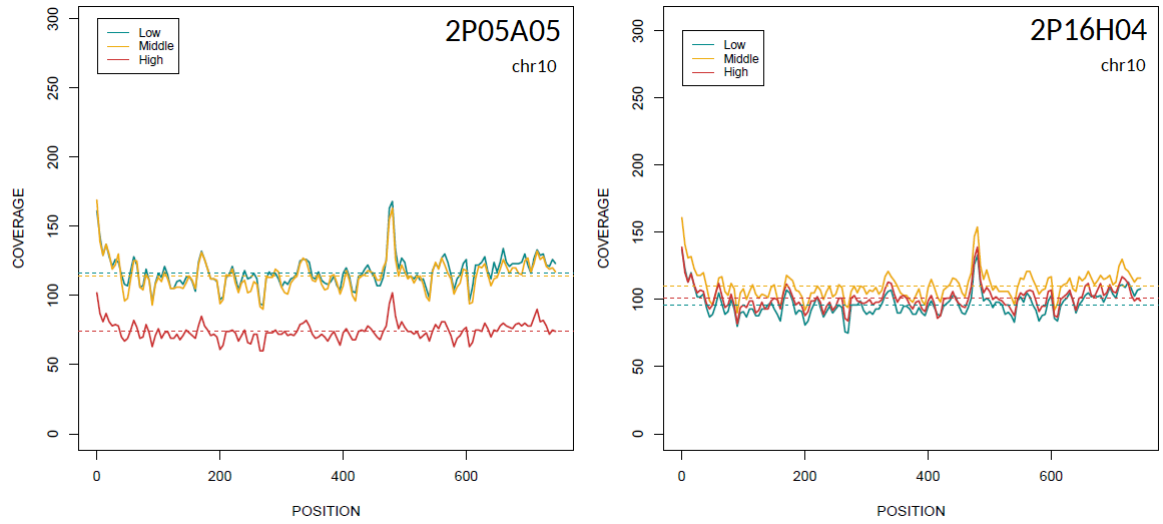


FIGURE 13 – Comparaison du taux de couverture entre deux segregants pour le chromosome 10

La couverture correspond au nombre de reads alignés en moyenne par position. Pour chaque segregant on a tracé la couverture de chaque bulk pour chaque position dans le génome. L'ensemble des bulks de tous les segregants fluctue en moyenne autour de 100. Ceci concorde avec les manipulations expérimentales puisque les quantités d'ADN de chaque bulk ont été dosées de façon à obtenir une couverture de 100. De plus, en théorie l'ADN de chaque bulk doit être présent dans les mêmes proportions et donc l'échantillon séquencé contenait les mêmes quantités d'ADN issus de chaque bulk. On observe toutefois

que ce n'est pas le cas pour 2 segregants. En effet, pour le segregant 2P05A05 par exemple la couverture du bulk high est plus faible (en moyenne 75) que la couverture des bulks low et middle (en moyenne 110). On en conclut que la quantité d'ADN du bulk high était plus faible que la quantité d'ADN des bulks low et middle et donc que le nombre de reads du bulk high était moins élevé (expliquant une couverture plus faible). Pour 3 segregants, les couvertures des bulks low, middle et high fluctuent toutes autour de 100 donc on en conclut que les quantités d'ADN pour chaque bulk étaient présentes en proportion égale.

On observe des fluctuations dans la couverture en fonction des positions et on voit également que ces fluctuations ne dépendent pas des bulks. Ces fluctuations sont dues au fait que les bases C et G mettent plus de temps à être séquencées lors de l'amplification par PCR. Les fragments contenant plus de C et G sont donc présents en plus faible proportions à la fin de la PCR. Lors de l'alignement des reads, les régions contenant des C et G sont donc moins couvertes.