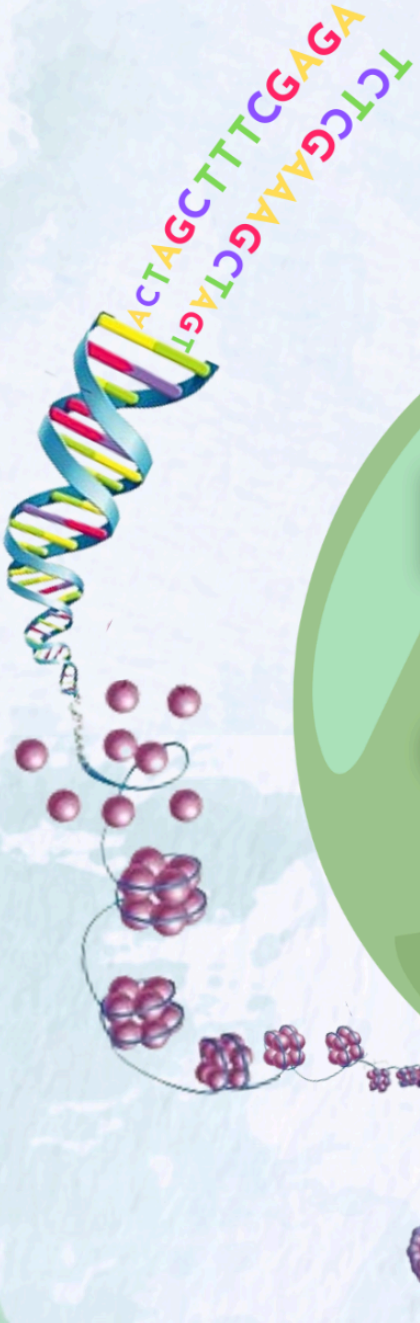
Analyse comparée de la diversité de loci de résistance aux maladies chez le melon

Marie Roynette

Marie ROYNETTE

IUT Sénart-Fontainebleau

BUT2
Génie Biologique
parcours
Sciences de l'Environnement et Écotechnologies (FI)



Rapport de stage

Analyse comparée de la diversité de loci de résistance aux maladies chez le melon

INRAE, unité EPGV

2 rue Gaston Crémieux, Évry

Responsable de stage : Patricia FAIVRE RAMPANT

Du 02/04/2024
au 31/07/2024

UPEC
UNIVERSITÉ PARIS-EST CRÉTEIL

IUT SENART FONTAINEBLEAU

INRAE

EPGV
Etude du Polymorphisme des Génomes Végétaux

Remerciements

Je remercie tout d'abord ma maître de stage Patricia FAIVRE RAMPANT, directrice de l'unité EPGV, pour sa confiance lors de mon recrutement et tout au long de mon stage. Je remercie également Damien HINSINGER, ingénieur de recherche, ainsi qu'Aurélie CANAGUIER, bio-analyste. M'ayant tous trois assistée au cours de mon stage, je les remercie pour leur accompagnement constant, le partage de leurs connaissances ainsi que pour le suivi et la relecture de ce rapport.

Je tiens à adresser ma considération pour le reste de l'équipe de laboratoire, Aurélie BERARD, responsable et ingénieure d'étude, et les techniciens Isabelle LE-CLAINCHE et Romain MESNIL. J'aimerai également mentionner Raphaël MINGUELLA, stagiaire de M2 à l'EPGV, qui a eu la gentillesse de me partager 2 de ses précédents rapports de stage et projet en tant que modèles pour ma rédaction.
Je remercie bien sûr Javier BELINCHON-MORENO, doctorant de la thèse qui a été le support de mon stage et l'ensemble de l'équipe EPGV pour le partage de toutes ces connaissances et compétences mais aussi pour leur accueil et leur convivialité/bienveillance tout au long de mon stage.

Je souhaite aussi adresser mes remerciements à l'équipe pédagogique de l'IUT de Sénart-Fontainebleau pour la qualité de l'enseignement et leur soutien dans le cadre de ce stage de BUT 2, ainsi qu'à Tharaniya TAMBOSCO pour son accompagnement en tant que responsable des stages à l'IUT.

Enfin, je voudrai exprimer ma reconnaissance envers Florence RIBIÈRE, amie et technicienne au CNRGH, qui a eu la gentillesse de transmettre ma candidature à Patricia FAIVRE RAMPANT, m'ayant ainsi permis une première prise de contact avec l'EPGV qui m'a grandement aidée pour l'obtention de ce stage.

Sommaire

Introduction

Dans le cadre de ce stage de 2ème année de BUT Génie Biologique parcours Sciences de l'Environnement et Écotechnologies, j'ai intégré l'unité EPGV (Étude du Polymorphisme des Génomes Végétaux) de l'Institut National de Recherche pour l'Agronomie, l'Alimentation et l'Environnement (INRAE) localisée à Evry. Mon travail repose sur la thèse en génomique végétale de Javier BELINCHON MORENO, associant l'EPGV pour les données de séquençage, l'Université d'Avignon et l'unité GAFL (Génétique et Amélioration des Fruits et Légumes) d'INRAE à Avignon pour le matériel végétal et l'expertise agronomique.

La génomique végétale est un domaine de recherche visant à caractériser et comprendre les mécanismes génétiques des plantes et leurs adaptations à leur environnement. Les études s'y rapportant peuvent conduire à des sélections variétales, ayant de nombreux objectifs tels qu'améliorer les rendements de culture, diminuer l'utilisation des pesticides, garantir la sécurité alimentaire ou encore préserver la biodiversité. Au cours de sa thèse, Javier BELINCHON MORENO étudie la variabilité intraspécifique d'une famille de gènes de résistance (NLR) chez le melon. L'analyse des variations génétiques ou polymorphismes, au sein d'une population, permet d'explorer la diversité intraspécifique et d'approfondir les connaissances sur ces gènes NLR dont la structure et les fonctions restent encore mal connues.

Mon objectif est de reconstruire 6 régions génomiques contenant des gènes NLR incomplets pour 144 variétés de melon puis d'étudier leur structure et les polymorphismes intraspécifiques. Pour cela, je dois d'abord définir la démarche de reconstruction des régions d'intérêt puis l'automatiser pour l'analyse des 144 variétés. Ce travail sera conduit à l'aide d'un logiciel de bio-informatique spécifique pour l'étude génomique. Je pourrai ensuite étudier les structures et polymorphismes des séquences finales grâce à des plateformes génomiques en ligne.

Mon stage repose donc sur la problématique suivante :
Afin d'étudier le polymorphisme de gènes de résistance dans 6 régions d'intérêt du melon, comment reconstruire ces séquences à l'aide d'un logiciel informatique spécifique, et automatiser cette démarche ?

I. Présentation de l'organisme d'accueil et contexte du stage

I.1 Présentation de l'organisme d'accueil

INRAE, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, est un établissement public de recherche scientifique en Sciences de la Terre et du Vivant, réparti sur 18 centres en France métropolitaine et outre-mer. INRAE est divisé en 14 départements de recherches spécialisées dans différents domaines[1], dont le département Biologie et Amélioration des Plantes (BAP) auquel j'étais rattachée pour mon stage.

Le département BAP mène des projets dans le domaine des sciences du végétal dans le but d'apporter des connaissances sur le fonctionnement des plantes en interaction avec leur environnement biotique et abiotique pour comprendre et optimiser le fonctionnement des agrosystèmes. Dans ce département BAP, on retrouve l'unité EPGV dans laquelle j'ai effectué mon stage.

Cette unité d'Étude du Polymorphisme des Génomes Végétaux est une plateforme de génomique INRAE rattachée au Centre Ile-de-France Versailles-Saclay mais localisée à Évry sur le site des 2 grands centres nationaux de séquençage, le Génoscope et le CNRGH (composants de l'Institut de biologie François Jacob du CEA). L'EPGV répond aux projets en génomique végétale et environnementale pour l'étude de polymorphismes ponctuels (SNP) et des variations structurales. Actuellement, l'EPGV travaille sur divers projets comme le développement d'outils génomiques (génotypage) pour la description de collections de lentilles, de lin, maïs, melon, etc.

I.2 Contexte génomique et modèle d'étude

Pour ce stage, j'ai eu l'occasion de m'immerger au cœur de la génomique, discipline de la biologie moléculaire visant à étudier la structure et le fonctionnement d'un organisme, organe, cellule, etc. à l'échelle du génome. Le génome est l'ensemble du matériel génétique d'une espèce contenu dans l'ADN. Les constituants élémentaires de l'ADN portant l'information génétique sont les nucléotides, organisés en séquences. Parmi ces séquences, certaines sont codantes, formant ainsi les gènes, éléments d'information héréditaire qui permettent la synthèse de différentes protéines dont l'expression affecte les caractères de l'organisme. Les caractères visiblement exprimés constituent le phénotype de l'individu. Des variations de quelques nucléotides, appelés polymorphismes, peuvent être observées entre plusieurs variétés d'une même espèce et faire varier les phénotypes si elles se situent notamment dans la séquence d'un gène, et ainsi altérer sa fonction. Dans le génome, les séquences d'ADN non codantes jouent aussi un rôle déterminant, réparties en tant qu'éléments répétés, introns, ou séquences impliquées dans l'organisation et la maintenance du génome ou encore la régulation de certains processus biologiques[2].

L'étude du génome des plantes permet notamment d'en apprendre davantage sur les mécanismes de leur immunité, induite par les gènes de résistance, aussi appelés gènes R. La plus grande famille de gènes de résistance aux pathogènes et ravageurs est formée par les gènes de type NBS LRR (*Nucleotide-Binding Site Leucine-Rich-Repeat*) ou plus communément NLR. L'ensemble de ces gènes parmi toutes les variétés d'une espèce est appelé NLRome de référence. Ces gènes ont en commun plusieurs domaines protéiques[3], c'est-à-dire une portion particulière du gène codant pour certaines fonctions spécifiques de la protéine complète (ici la spécificité de résistance). Les gènes NBS-LRR possèdent un domaine LRR, composé de répétitions riches en leucines, permettant la reconnaissance de l'agent pathogène. Ensuite, un domaine NBS

(« nucleotide-binding site ») de liaison de nucléotides induit un changement de conformation de la protéine activant ainsi la fonction de résistance. Enfin, un domaine TIR ou CC en position N-terminal jouent un rôle dans l'interaction protéine-protéine et hôte-agent pathogène[4].
Malgré les connaissances accumulées, le rôle spécifique de chaque gène NLR reste largement méconnu, en particulier dans les résistances quantitatives[5], c'est-à-dire les résistances polygéniques impliquant une résistance partielle, contrairement à la résistance qualitative qui s'exprime de façon binaire (résistant ou sensible). Cette méconnaissance des gènes NLR s'explique en partie par la complexité de ces régions[6]. En effet, les gènes NLR présentent un grand nombre de répétitions arrangées en tandem, formant ainsi des clusters, c'est-à-dire une succession de plusieurs gènes rapprochés. Dans ces clusters, on retrouve également de nombreux éléments transposables (ET), des séquences observables à plusieurs endroits du génome. La faible expression de ces gènes (expression basale) rend d'autant plus difficile leur identification. La structure et l'organisation complexe de ces gènes a rendu compliqué le processus de séquençage/assemblage/annotation avec les technologies courtes lectures. L'accessibilité du séquençage longues lectures a donc permis d'obtenir une qualité de séquençage compatible avec des études plus approfondies des gènes NLR.

Le modèle choisi dans le cadre de l'étude est le melon. De son nom scientifique *Cucumis melo* L., le melon est une angiosperme (ou « plante à fleurs ») de la famille des Cucurbitacées, décrite pour la première fois par Carl von Linné (abrégé en L.). C'est une plante allogame qui est donc fécondée par le pollen d'une autre plante de la même variété ou espèce, principalement par les abeilles. Le melon est une plante maraîchère venue d'Afrique tropicale et arrivée en France en 1495[7]. Le melon est une culture horticole économiquement importante avec 28,62 millions de tonnes produites à l'échelle mondiale en 2021[8] . Il existe près de 1000 variétés[9], caractérisées par leur teneur en sucre, acidité, couleur intérieure et extérieure, etc[10]. Le melon est victime de nombreux ravageurs tels que des bactéries (*Acidovorax avenae*, *Erwinia tracheiphila*…), des oomycètes et des fungi (*Fusarium oxysporum f.sp. melonis*, *Pseudoperonospora cuben*…) ou encore des virus (genres *Cucumovirus*, *Potyvirus*, *Begomovirus*…[11])[12] et [13]. Ces ravageurs peuvent induire de lourdes pertes économiques en détruisant les cultures.
À l'échelle génomique, le melon compte parmi les plus petits génomes végétaux des espèces cultivées[14] (375Mb à 400Mb). Son génome est organisé en 12 chromosomes. C'est une espèce diploïde (2n = 24) et principalement homozygote, c'est-à-dire que les 2 allèles de chaque gène sont identiques entre eux, elle ne possède donc qu'une seule version par gène. C'est également l'un des génomes contenant le moins de gènes de résistance, avec 84 gènes R dénombrés[15]. Les gènes NLR chez le melon couvrent ~1% de son génome.
En comparaison, la vigne possède un génome de taille similaire mais avec environ 400 gènes hétérozygotes de résistance. Ainsi, le melon est particulièrement adapté à la caractérisation complète du NLRome dans un panel de diversité du fait de la petite taille de son génome, du faible nombre de NLR qu'il contient et de la relative simplicité d'étude dans un cadre d'homozygotie.

I.3 Contexte du sujet (de stage)

Le travail réalisé au cours de mon stage est intégré à la thèse *Diversité génétique et fonctionnelle du NLRome/résistome chez le melon* de Javier BELINCHON-MORENO dont l'objectif est de comprendre le déterminisme génétique de la résistance du melon à différents agents pathogènes. Afin d'approfondir les connaissances sur les gènes NLR et leur rôle dans la résistance aux agents pathogènes et maladies végétales, cette thèse fixe l'enjeu scientifique de décrire le NLRome de *Cucumis melo*, caractériser ces gènes et leur diversité génétique, et enfin de comprendre leurs rôles dans l'expression de l'immunité face un large cortège de ravageurs et d'agents pathogènes.

Comme expliqué dans l'article[15] présentant son travail de thèse, des melons cultivés à Avignon ont d'abord été phénotypés, c'est-à-dire décrits par leurs caractères visibles, afin de répertorier l'expression des résistances. En parallèle, les génomes des mêmes variétés ont été séquencés par l'EPGV à Évry pour l'étude de polymorphismes SNP (génotypage). Des liens statistiques reliant phénotypes et génotypes peuvent être établis et constituent la base de la génétique d'association ou GWAS (genome-wide association study). Elle permet l'étude entre variabilité phénotypique et variabilité du NLRome sur les 144 variétés.

La méthode de séquençage utilisée pour le travail de thèse est également une preuve de concept de la caractérisation génomique de séquences ciblées et complexes par séquençage longues lectures, le NAS (cf. *II. Matériel et méthodes*).

Dans le cadre d'un précédent projet, un assemblage *de novo* du génome de la variété Anso77 a été réalisé, c'est-à-dire que tous ses fragments d'ADN séquencés ont été mis bout à bout par correspondance de nucléotides afin d'obtenir une première version de sa séquence génomique[16]. Ce génome a également été annoté, ce qui signifie que les différents éléments le composant (gènes, séquences intergéniques, éléments répétés, etc) ont été identifiés et bien définis, permettant une connaissance précise de la structure du génome. Ainsi, Anso77 était déjà bien connue par les chercheurs d'Avignon du point de vue génomique et agronomique. Cette variété a donc naturellement été choisie comme référence pour la comparaison des gènes de résistance avec les autres variétés.

Quinze régions d'intérêt ont été identifiées dans le génome d'Anso77 d'après l'analyse des données produites par NLGenomeSweeper, un outil pour l'identification des domaines des gènes de résistance NBS-LRR à l'échelle du génome[17]. Chaque région contient un ou plusieurs gène(s) d'intérêt entouré(s) de 20kb de séquence de part et d'autre (±20kb) afin d'étudier le contexte génomique tels que les régions régulatrices, s'il est isolé ou se trouve dans un cluster, etc. Ces régions ont été extraites et une annotation expertisée a été réalisée (promoteurs, introns, exons, codons d'initiation et stop, séquences régulatrices, etc). Les séquences de ces 15 régions chez Anso77 sont ensuite enregistrées dans un fichier au format fasta (fichier contenant une liste de séquences nucléotidiques et leurs noms).

Deux types de séquençage (cf. *II.2 Matériel génomique*) ont respectivement permis d'obtenir des séquences longues lectures (>1kb) et des séquences courtes lectures (≈150 bases). Tout d'abord pour chacune des 144 autres variétés, leurs gènes NLR ont été ciblés pour un séquençage en longues lectures NAS. Cette méthode se base sur l'homologie de séquence entre les régions d'intérêt de la référence et les séquences des variétés d'intérêt. L'assemblage de ces lectures a permis de reconstituer les séquences complètes de chaque région des 144 variétés. Afin de réaliser une analyse fine du polymorphisme du NLRome, un séquençage tout génome en courtes lectures sans assemblage a ensuite été réalisé.

Après l'annotation identifiée par NLGenomeSweeper, la comparaison génomique avec d'autres génomes de melons a permis de mettre en évidence que 6 gènes avaient été oubliés. N'étant de ce fait pas répertoriés dans le fichier fasta de référence dès le début du séquençage NAS, leurs séquences n'ont pas pu être sélectionnées chez les 144 autres variétés. Il serait trop long et trop coûteux de refaire un séquençage NAS en incluant les séquences de ces 6 gènes dans le fichier fasta de référence. En revanche, ces séquences sont présentes parmi les courtes lectures et potentiellement dans certaines longues lectures NAS également (cf. *II.2 Matériel génomique*). Ainsi, toutes les lectures NAS à disposition ainsi que les courtes lectures ont été récupérées pour essayer de reconstruire les régions de ces 6 gènes manquants.

C’est dans ce contexte qu’intervient mon travail de stage.
L’objectif est de reconstituer, pour les 144 variétés, les 6 régions génomiques manquantes positionnées chez Anso77, puis de comparer ces séquences entre elles pour identifier les polymorphismes. Mon travail consiste à étudier comment atteindre cet objectif final sur 8 variétés et à créer un workflow (ou suite d’étapes liées) afin de les exécuter de manière automatique sur les 135 autres variétés. C’est donc un travail de bio-analyste.

II. Matériels et méthodes

II.1 Matériel végétal

Mon travail de stage a pour objectif d’étudier ces 6 gènes de résistance pour 144 variétés de melon. Au cours des 2 premiers mois de stage, j’ai travaillé avec 9 variétés, Anso77 en témoin d’étalonnage et 8 autres, qui constituent le jeu test (***Tableau 1***).

Tableau 1 : Description des 9 variétés de melon du jeu test

Variété	Origine (classement pays producteurs[18])	Groupe botanique
ANSO77 (témoin positif)	Espagne (8e)	Inodorus
MOQUATRE	France (16e)	Cantalupensis
EARLS88632	Japon (22e)	Reticulatus
K2005	Chine (1er)	Makuwa
PI116479	Inde (3e)	(Non connu)
HSD753	Soudan (45e)	Tibish et Seinat
SHENDI		Flexuosus
FQUS	Tunisie (26e)	
CUM413		

Les 135 autres variétés à étudier sont originaires de pays variés, majoritairement répartis en Europe méridionale, Afrique du Nord, Moyen-Orient et Asie du Sud représentant 17 ou 18 groupes botaniques (***Annexe 1***).

II.2 Matériel génomique

Mon stage étant focalisé sur l'aspect bio-informatique, je n'ai pas participé à des manipulations en laboratoire pour l'obtention des séquences. Toutes les données dont j'avais besoin pour les analyses et comparaisons ont été mises à ma disposition car les séquençages avaient déjà été réalisés. Il est cependant important de connaître ces différents séquençages effectués afin de comprendre les données sur lesquelles j'ai travaillé.

Il existe plusieurs types de séquençage en fonction des objectifs et des besoins d'analyse qui suivent. La ***Figure 1*** schématise rapidement les types de séquençage que j'ai eu l'occasion d'exploiter ou de discuter au cours de mon stage.

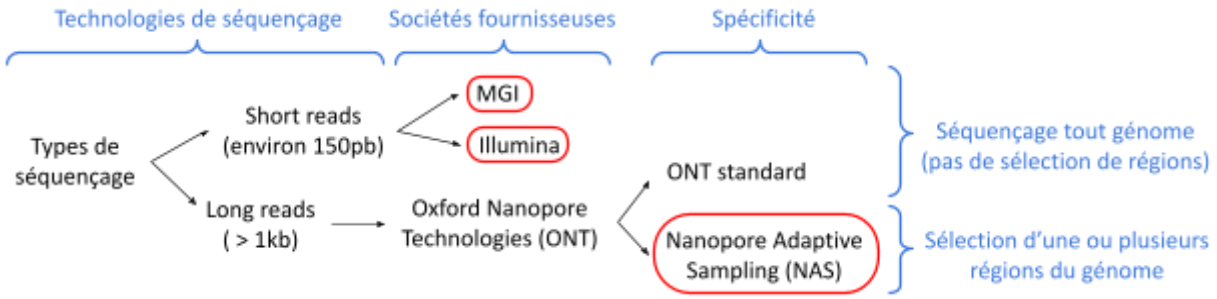


Figure 1 : Schéma des différents séquençages réalisés pour l'étude du NLRome de melon
Encadré en rouge, techniques ayant fourni les données analysées lors de mon stage

Pour mon travail de stage, j'ai donc exploité les 2 types de données de séquences suivantes :

➢ Les courtes lectures en paired-end issues de séquençages MGI ou Illumina

Les séquençages MGI et Illumina produisent des courtes lectures, c'est-à-dire des lectures d'ADN d'une taille jusqu'à 150 bases (***Figure 2***). Ces types de séquençage sont utilisés pour couvrir tout le génome de chaque variété. Le séquençage courtes lectures peut être réalisé en single-end, c'est-à-dire que les 150 premières bases de fragments d'ADN sont séquencés dans un sens unique. Cependant dans le cadre de la thèse, le séquençage a été réalisé en paired-end, ce qui consiste à séquencer 150 bases à chaque extrémité de fragments d'ADN d'environ 500 bases. Deux lectures en paire par fragment d'ADN sont produites en conservant l'information qu'elles appartiennent au même fragment d'ADN, permettant de déduire la distance approximative qui les sépare. Cela permet par la suite de confirmer le bon alignement de ces courtes lectures et d'être "plus stringent" sur leur position dans le génome. Parmi toutes ces courtes lectures, j'ai disposé de celles s'alignant sur les régions d'intérêt, extraites par Javier BELINCHON-MORENO.

Figure 2 : Exemple de fichier courtes lectures (Anso77 Chr10:15670269-15670716) vu sous Geneious

➢ Les longues lectures ONT approche NAS

Le séquençage ONT (Oxford Nanopore Technologie) utilise des nanopores, c'est-à-dire des protéines lectrices d'ADN, pour séquencer des fragments d'ADN de 1 à 50 kb, appelées longues lectures (***Figure 3***) . Dans le séquençage ONT, on distingue le séquençage ONT standard[19] (***Annexe 2***) qui permet de séquencer en longues lectures l'entièreté d'un génome sans sélection de région, ainsi que le séquençage ONT *Nanopore Adaptive Sampling* ou NAS[20] (***Annexe 3***), utilisé dans le cas de cette étude. Ce dernier permet le même type de séquençage mais en sélectionnant une région particulière du génome en comparant en temps réel les molécules d'ADN séquencées à un fichier de référence au format fasta. Tout fragment d'ADN est séquencé sur environ 500 bases le temps que la séquence soit identifiée et comparée. Si elle correspond au fichier fasta, le fragment d'ADN appartient à une des régions d'intérêt et est donc entièrement séquencé. En revanche, si elle ne correspond pas, les 500 premières bases séquencées sont conservées dans le fichier de séquence mais la molécule est rejetée.

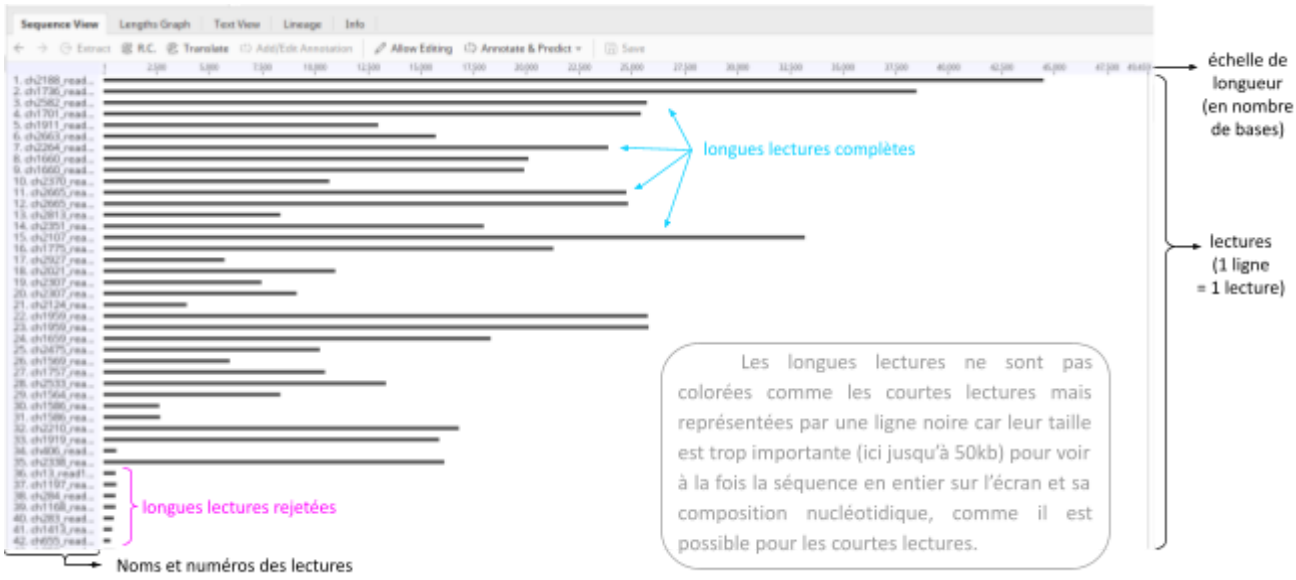


Figure 3 : Exemple de fichier longues lectures NAS (Anso77 Chr06:27029991-27032305) vu sous Geneious

Pour mon travail, j'ai exploité à la fois les longues lectures complètes et rejetées. Sachant que les 6 gènes recherchés sont de la même famille que ceux étudiés précédemment dans le cadre de la thèse, ils contiennent le même domaine de résistance dans leur séquence. Dans la mesure où le NAS se base entre autres sur ce domaine pour la comparaison des séquences, l'hypothèse de départ est qu'une partie des régions d'intérêt manquées devrait avoir été séquencée en longues lectures complètes. Les lectures couvrant les régions d'intérêt devraient également se trouver parmi les lectures rejetées.

Afin d'exploiter ces données, je disposais, en plus des lectures des 144 variétés, des séquences de référence d'Anso77 correspondant aux 6 gènes ±20kb décrits dans le ***Tableau 2*** et la ***Figure 4*** .

Tableau 2 : Caractéristiques des 6 gènes étudiés

Numéro du gène	Nom du gène	Position du gène d'intérêt (locus)	Région en nom abrégé	Nombre de bases de la séquence du gène ou du domaine
1	MELO3C016529	Chr06:27029991-27032305	Chr6a	2 315
2	MELO3C014062	Chr06:32491132-32494301	Chr6b	3 170
3	MELO3C005506	Chr09:20704787-20706019	Chr9a	1 233
4	MELO3C005690	Chr09:22204398-22209165	Chr9b	4 768
5	MELO3C022580	Chr10:15670269-15670716	Chr10	448
6	MELO3C002574	Chr12:22182931-22183739	Chr12	809

Figure 4 : Séquences de référence d'Anso77 des 6 régions d'intérêt (gène en bleu ± 20kb)

Enfin, j'avais également accès au génome complet d'Anso77 organisé en chromosomes, permettant de tester quelques assemblages ou comparaison suite à la reconstruction complète des régions d'intérêt des nouvelles variétés.

Pour résumer, les données utilisées pour mon travail de stage sont une sélection de courtes lectures en paired-end d'environ 2x150 bases qui couvrent les régions d'intérêt ainsi que des longues lectures ONT NAS de ~500 bases à plusieurs kb. Toutes ces lectures couvrent les 6 régions d'intérêt sur environ 40kb (gène ±20kb). Je dispose pour chaque région et variété de 2 fichiers au format fasta et de la séquence de référence d'Anso77 pour ces 6 régions d'intérêt.

II.3 Principe conceptuel de la démarche

Pour reconstruire des séquences génomiques à partir de lectures d'ADN, 2 concepts sont importants à définir.

Tout d'abord, l'alignement est l'action de superposer des lectures sur une séquence de référence en fonction de leurs similarités nucléotidiques. Il donne un aperçu global de toutes les lectures qui correspondent à la séquence de référence, et met en évidence la couverture, ou profondeur de séquence, représentant le nombre de lectures couvrant 1 position de la référence. Voir des exemples d'alignement en ***Figures 5.a et 5.b*** .

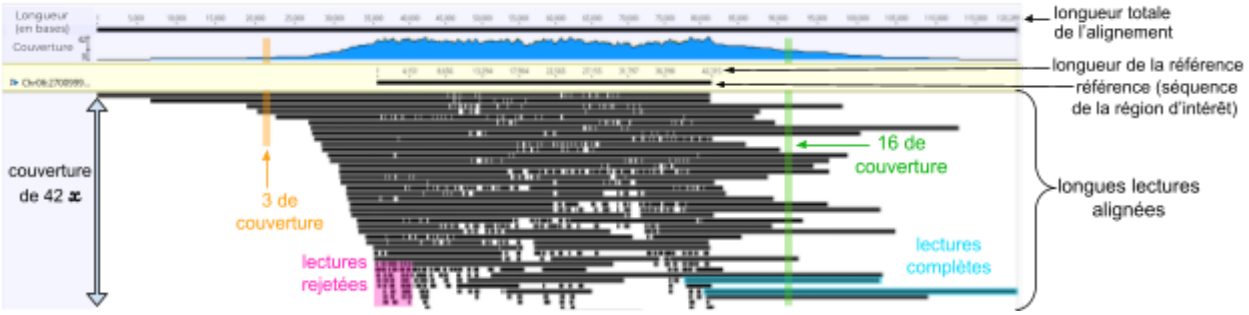


Figure 5.a : Visualisation de l'alignement de longues lectures sur une séquence de référence sous Geneious Prime

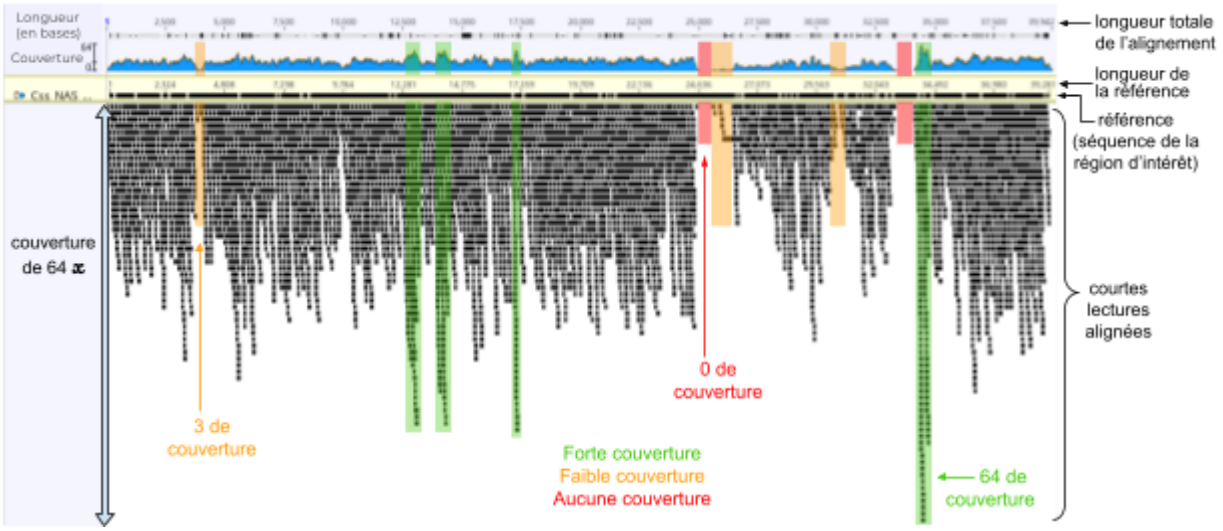


Figure 5.b : Visualisation de l'alignement de courtes lectures sur une séquence de référence sous Geneious Prime

À partir de l'alignement, un consensus peut être généré, ce qui correspond à établir une seule séquence à partir de toutes les lectures alignées en sélectionnant la base majoritaire sous chaque nucléotide de la référence. Ainsi, une seule séquence correspondant à la région d'intérêt de la nouvelle variété est créée. La ***Figure 6*** montre un exemple de formation du consensus à partir de l'alignement.

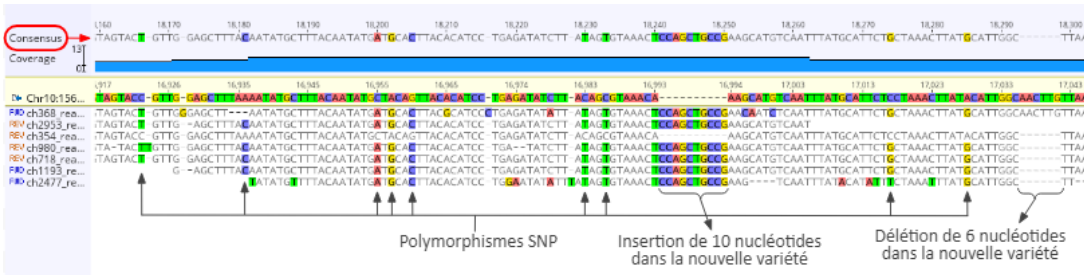


Figure 6 : Schéma d'un consensus à partir d'un alignement sous Geneious Prime

Enfin, le dotplot est un type de graphique statistique permettant de représenter le polymorphisme d'alignement entre 2 séquences. Il permet de visualiser les différences entre les 2 séquences, par exemple les insertions et délétions (***Figure 7*** dotplot 1) mais aussi les séquences répétées constituant les microsatellites (***Figure 7*** dotplot 2).

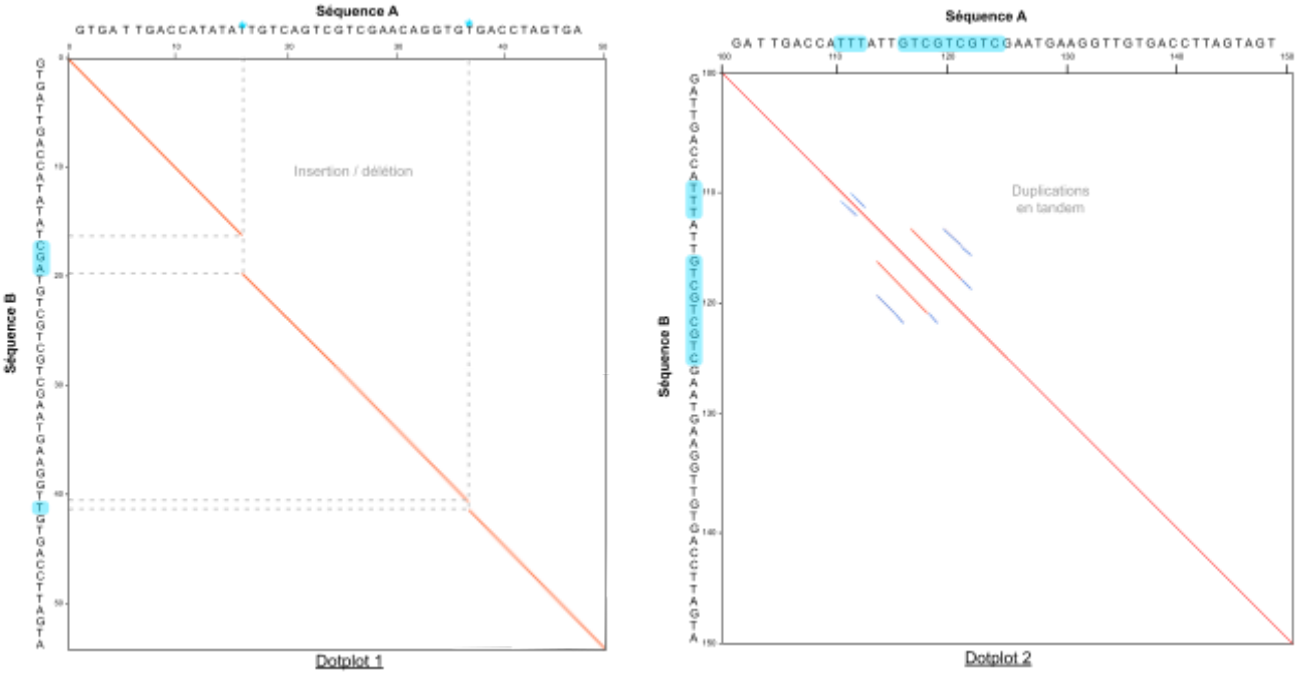


Figure 7: Schéma de dotplots de 2 régions des séquences A et B différentes (Dotplot 1) et similaires (Dotplot 2)

II.4 Description du logiciel Geneious Prime

Toutes mes analyses ont été conduites avec le logiciel Geneious Prime[21].
Geneious Prime est un logiciel de bio-informatique disponible pour l'analyse des données de séquences en génomique. Il peut être utile pour l'étude et la comparaison de génome d'un ou plusieurs individus, l'annotation de gènes, la recherche de polymorphismes et variations structurales, la construction d'arbres phylogénétiques… Geneious Prime permet également de comparer des données avec des données de référence provenant de banques mondiales telles que NCBI[22] et UniProt[23].

Avec le logiciel Geneious Prime, il est possible de créer des workflows, c'est-à-dire des programmes de tâches informatiques listant les étapes que l'on souhaite effectuer à partir d'une sélection de fichiers d'entrée afin qu'elles se déroulent automatiquement. Cette option est indispensable dans le cadre de mon travail dans la mesure où le but de l'étude est de traiter 6 régions sur chacune des 144 variétés. Sachant que chaque reconstitution et analyse demande plusieurs étapes, il ne serait pas optimal de réaliser cela manuellement.

II.5 Prise en main du logiciel et étude des paramètres

La première étape de mon travail était une phase de questionnement et de réflexions sur les bases de ma démarche. J'ai testé les différents paramètres et options disponibles sur Geneious Prime, ainsi que la méthode la plus adaptée pour obtenir les consensus pour chaque variété. L'objectif est de définir les fichiers d'entrée (courtes et/ou longues lectures), la combinaison de paramètres et la hiérarchie des étapes qui permettraient d'obtenir le consensus final le plus fiable pour chaque région de chaque variété.

Pour cela, j'ai d'abord étudié 2 variétés avec la référence Anso77. Tout d'abord, les données d'Anso77 ont servi de témoin positif. Étant de la même variété que la référence, les séquences obtenues devraient être identiques à celle de la référence, seuls les paramètres choisis peuvent induire une différence. En parallèle, j'ai travaillé avec Moquatre, une variété génétiquement éloignée d'Anso, ayant un groupe botanique et pays d'origine différent (cf ***Tableau 1*** dans *II.1 Matériel végétal*).

Sur Genious Prime, la modification des paramètres est accessible dans le module de l'alignement puis de la création du consensus. Pour cette première étude avec Anso77 et Moquatre, j'ai donc réalisé manuellement des alignements et leur consensus à partir des longues lectures puis des courtes lectures et enfin en combinant les deux. En parallèle, j'ai testé différents paramètres afin de repérer comment les résultats variaient. Ces tests aident donc à fixer les paramètres et identifier la meilleure démarche.

Paramètres d'alignement :

➢ Tout d'abord, il faut définir la qualité minimale souhaitée pour les bases des lectures. Cette qualité dépend du séquençage et correspond à la certitude que le séquenceur ait correctement identifié la base. Une fois l'alignement réalisé, cette qualité peut être indiquée lorsque la base ne correspond pas à la séquence de référence pour savoir s'il s'agit d'une potentielle erreur de séquençage ou d'un réel polymorphisme.

➢ Ensuite, la qualité de l'alignement dépend du taux de mismatchs autorisés, c'est-à-dire chaque position où la référence et la lecture sont différentes. Il peut s'agir d'un SNP, d'une délétion ou d'une erreur d'alignement de la lecture. Ce paramètre de Minimum Mismatch (MM) étant fixé à 30% par défaut, les tests ont permis de décider s'il fallait changer cette valeur, sachant que 30% de mismatch n'a pas le même impact selon s'il s'agit d'aligner des courtes lectures de 150 bases ou des longues lectures de 50kb.

➢ Pour finir avec les paramètres de qualité d'alignement, il faut se pencher sur la taille permise pour les ouvertures, c'est-à-dire lorsque la référence et les lectures ne correspondent pas sur une portion. La taille de gaps influe sur la possibilité de Geneious Prime de repérer les insertions, délétions et changements de structure. En effet, si la nouvelle variété présente une insertion d'une taille supérieure à la taille maximale d'ouvertures autorisée, l'insertion ne sera pas prise en compte. Si une taille trop basse peut être problématique dans la recherche de variations structurales, une taille d'ouverture trop élevée peut quant à elle permettre à des lectures n'appartenant pas à la région d'être incorporées dans l'alignement, ce qui n'est pas souhaité.

Ces paramètres pour l'alignement sur Geneious Prime, pouvant être modifiés et influant sur la séquence finale, sont présentés en ***Annexe 8***.

Paramètres de consensus :

➢ Premièrement, la couverture lors de l'alignement est déterminante dans la formation du consensus. Le nombre de lectures alignées sous une même position de la séquence de référence doit en effet être suffisant pour pouvoir établir un consensus fiable du nucléotide majoritaire à chaque position. Le manque de couverture perturbe la création du consensus (***Figure 8***). En effet, le consensus ne peut pas être généré et affiche alors soit des N (aucune base) soit intègre la séquence de référence à la place, en fonction du paramètre sélectionné.

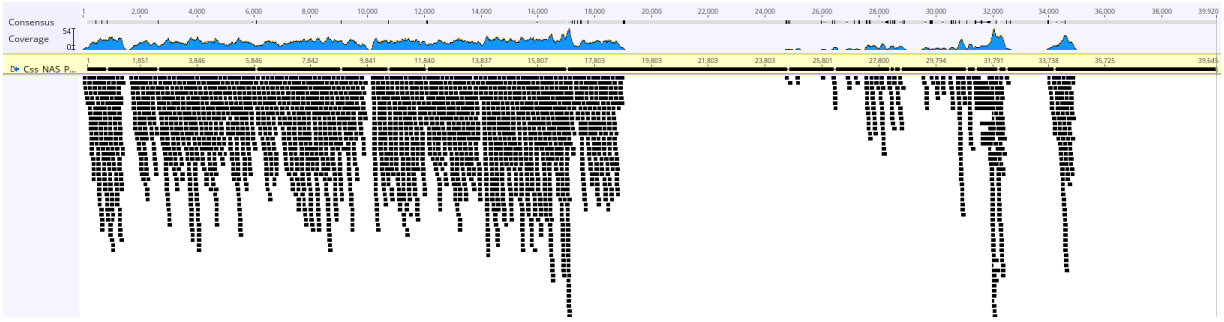


Figure 8 : Alignement courtes lectures montrant une absence de couverture

Afin de déterminer le paramètre optimal pour générer le consensus en cas d'absence de couverture, j'ai réalisé un arbre de décision (***Annexe 9***). Ce dernier présente tous les cas théoriquement observables en cas d'absence de couverture longues et/ou courtes lectures, les hypothèses de leurs significations et les résultats attendus pour différentes combinaisons de paramètres. Sur l'ensemble des cas, j'ai cherché à identifier la combinaison de paramètres permettant d'obtenir un consensus final le plus conforme possible à la variété d'intérêt et non à la référence, et minimiser la présence de N.

➢ Deuxièmement, il faut s'intéresser au seuil de consensus. Habituellement, le consensus est établi d'après la base majoritaire sans condition. Cependant, la majorité uniquement n'est pas toujours représentative et cela peut être problématique en cas d'égalité entre nucléotides. Le consensus généré n'étant pas décisif, le logiciel peut se référer au code IUPAC, détaillé en ***Annexe 7***. La présence de nucléotides dégénérées dans le consensus peut par la suite compromettre l'annotation et l'analyse des séquences.
Il est donc possible d'exiger un pourcentage de majorité, appelé seuil de consensus, ne retenant ainsi que la base atteignant ce pourcentage. Le seuil de consensus permet donc d'obtenir un consensus plus fiable, en choisissant la base qui domine largement, excluant toute hétérozygotie ou erreur d'alignement. Cependant, il peut parfois apporter des difficultés supplémentaires par rapport à la simple majorité. Prenons l'exemple d'une position qui présenterait une couverture de 10 lectures, avec 4 C et 6 A. En cas de seuil à 65%, aucune base ne serait retenue et Geneious Prime se référerait alors au code IUPAC pour créer le consensus, affichant ainsi "M", tandis qu'avec un seuil à 50%, il n'y aurait pas eu de difficulté car la base A dépasse ce pourcentage de majorité.
Parmi toutes les possibilités pour le seuil de consensus, l'objectif est que les séquences finales ne contiennent, dans l'idéal, pas de lettres du code IUPAC autres que A, T, G, C et N (le moins possible pour cette dernière).

➢ Enfin, il faut s'assurer que les lectures sélectionnées pour créer le consensus depuis l'alignement correspondent bien à la région de référence d'Anso77. La possibilité d'ignorer les lectures s'alignant potentiellement à d'autres endroits du génome peut permettre de restreindre l'analyse à la région d'intéret uniquement.

L'***Annexe 10*** montre la fenêtre de modification de paramètres pour la création du consensus sur Geneious Prime.

Une première phase d'analyse des consensus obtenus m'a donc permis d'observer les résultats des tests pour tous ces paramètres et de me faire une idée de la meilleure démarche à suivre. Pour cela, j'ai observé un à un les alignements et consensus, et réalisé des dotplots.
Ces tests sont aussi l'occasion de vérifier que les capacités du logiciel correspondent bien aux objectifs de l'étude, à travers son comportement vis-à-vis des insertion/délétions et des motifs répétés.

II.6 Construction du workflow

Une fois m'être fait une première idée de la méthode grâce à la phase de tests, j'ai commencé à mettre en place le workflow. Cette deuxième phase a pour objectif de pouvoir générer les consensus des 144 variétés de manière automatique, en suivant les paramètres définis préalablement pour obtenir les meilleurs consensus.

J'ai d'abord commencé par réaliser des workflows très simples composés de 2 ou 3 commandes. Dans la mesure où la démarche complète pour créer les consensus finaux (cf *III.1 Démarche de reconstruction*) est composée de plusieurs étapes, il m'a fallu explorer les options de Geneious Prime pour agencer les différentes commandes et options. Je me suis inspirée des workflows de base du logiciel, servant justement de modèles et de test.
Une fois familiarisée avec le principe, j'ai ajouté davantage de commandes et réalisé des workflows plus complexes et adaptés à mon analyse. J'ai donc procédé commande par commande, en testant chaque nouvel ajout. J'ai également réparti des étapes de sauvegarde stratégiques afin d'identifier la commande qui pose problème en fonction du dernier fichier enregistré, en cas de non-aboutissement du workflow, .

Lors de la réalisation du workflow, j'ai rencontré quelques difficultés concernant l'agencement des étapes nécessaires ou bien des spécificités des fonctionnalités proposées. Ne parvenant pas à trouver de solution pour atteindre mon objectif, j'ai contacté par mail le service support de Geneious Prime (**Annexe 11**). Je les ai d'abord sollicités pour lier 2 étapes majeures du workflow afin de fluidifier la démarche, puis pour réduire la proportion de code IUPAC indésirable dans les consensus finaux (adaptation des paramètres d'alignement et consensus). Leur disponibilité sur plusieurs jours et nos échanges (en anglais) leur ont permis de suivre l'avancée de ma problématique de départ et de me guider à chaque blocage et incompréhension.

II.7 Exploitation des séquences consensus finales

L'objectif de cette étude est de caractériser la variabilité génétique intraspécifique des gènes NLR, leur structure et rôle dans la résistance aux maladies et agents pathogènes chez les plantes. Pour cela, j'ai extrait, sous forme de fichiers fasta, les consensus finaux des régions d'intérêt afin de les annoter et étudier les polymorphismes. L'annotation sur Genious Prime est peu ergonomique du point de vue de l'affichage des résultats et limitée en termes de bases de données internationales. J'ai donc opté pour la plateforme en ligne BLAST (Basic Local Alignment Search Tool) [24] servant à comparer des séquences nucléotidiques ou protéiques à des bases internationales, notamment le NCBI. Le BLASTx permet d'identifier dans une séquence nucléotidique, les zones qui correspondent à des protéines exprimées et connues dans les bases de données.

Pour le moment, je me suis penchée sur les 2 premières variétés étudiées lors de la phase de tests, Anso77 et Moquatre. Je me suis initiée à BLASTx avec le consensus de la région Chr6a de Moquatre pour la prédiction de gènes de résistance. Les résultats de BLASTx me permettront de caractériser la protéine codée par l'une des séquences nucléotidiques reconstruites, ce qui validerait la présence du gène de résistance. Sachant que le gène d'intérêt est entouré de ±20kb, j'ai ciblé les recherches BLASTx entre les 19.000^{e} et 25.000^{e} bases de la séquence.

III. Résultats et discussions

III.1 Démarche de reconstruction

Le ***Tableau 3*** résume les caractéristiques des consensus obtenus pour la région Chr6a avec les 4 scénarios envisagés lors de la première phase de tests (longues lectures uniquement, courtes lectures uniquement, longues et courtes lectures simultanément, longues puis courtes lectures) avec les paramètres par défaut de Geneious Prime.

Tableau 3 : Résultats des consensus de Anso77 et Moquatre de la région Chr6a pour les 3 scénarios d'alignement avec les paramètres par défaut de Geneious Prime

Variété	Caractéristiques	Types de lectures utilisées pour générer les alignements et consensus			
		Longues lectures	Courtes lectures	Longues et courtes lectures	Longues puis courtes lectures
Anso77	Taille consensus	42,309 bases	42,316 bases	42,317 bases	42,336 bases
	Code IUPAC	2R, 1K, 1Y	1N	1N, 1R, 1S, 1W	1N, 1K, 1W, 4S
	Couverture	$24 < \boldsymbol{x} < 161$ Moyenne = 37,5	$4 < \boldsymbol{x} < 74$ Moyenne = 18,2	$34 < \boldsymbol{x} < 161$ Moyenne = 54,2	$24 < \boldsymbol{x} < 161$ puis $2 < \boldsymbol{x} < 74$ Moyenne = 37,5 puis 18,2
Moquatre	Taille consensus	42,222 bases	42,266 bases	42,233 bases	42,215 bases
	Code IUPAC	24N, 3K, 2R, 2S, 1Y	1Y	9N, 1S, 1R, 1M	22N, 1M, 2W
	Couverture	$3 < \boldsymbol{x} < 177$ Moyenne = 13,3	$8 < \boldsymbol{x} < 282$ Moyenne = 23,8	$15 < \boldsymbol{x} < 329$ Moyenne = 34,1	$3 < \boldsymbol{x} < 177$ puis $0 < \boldsymbol{x} < 287$ Moyenne = 13,3 puis 24,4

NB : Taille référence Anso77 de la région Chr6a) = 42,315 bases

Dans tous les cas, Anso77 est stable en termes de taille de consensus par rapport à la référence. Quant à Moquatre, sa taille de consensus diffère de celles d'Anso77 et de la référence, traduisant un polymorphisme de taille entre Anso77 et Moquatre pour la région Chr6a.
D'après le ***Tableau 3***, on remarque la présence de code IUPAC dégénéré dans les consensus finaux des 2 variétés quelque soit le scénario de reconstruction choisi. Il est cependant plus important et variable pour Moquatre que pour Anso77 suggérant un polymorphisme nucléotidique entre les 2 variétés et un effet du scénario sur la reconstruction du consensus de Moquatre. Voici l'analyse des 4 scénarios :

1. L'alignement avec les **longues lectures** uniquement permet d'obtenir un consensus complet car les lectures couvrent entièrement la région d'intérêt à elles seules et le nombre de lettres du code IUPAC dégénéré est raisonnable. Cependant deux problèmes se posent. Bien qu'il y ait une bonne couverture moyenne sur toute la région d'intérêt, elle n'est parfois pas suffisante pour établir un consensus fiable (par exemple avec 3 de couverture). En général, une couverture de $5\boldsymbol{x}$ minimum est attendue pour valider le consensus. De plus, le séquençage NAS présente en moyenne 1-3% d'erreur de lecture des nucléotides. Ces deux critères de couverture insuffisante ou variable et d'erreurs de séquençage m'amène à la conclusion que les longues lectures seules ne suffisent pas pour un alignement et consensus de qualité.

2. L'alignement avec les **courtes lectures** uniquement permet d'obtenir un consensus complet. Les courtes lectures couvrent entièrement la région d'intérêt avec une couverture suffisante. De plus, par nature, elles présentent environ 0,1% d'erreur de lecture des nucléotides. Cela se remarque par la plus faible présence de code IUPAC et de polymorphismes visible par l'alignement. Bien que ce type de lectures semble optimal pour reconstruire un consensus, les tests effectués avec les courtes lectures seules mettent en évidence des alignements multiples aux endroits où on observe un pic de couverture. Par exemple, si la couverture moyenne est de 20𝒙 sur tout l'alignement et de 40𝒙 localement, cela peut signifier que la moitié des lectures proviennent d'un autre endroit du génome mais s'alignent ici car elles correspondent suffisamment à la séquence de référence. Ainsi, cet alignement ne correspond pas exclusivement à cette région et n'est donc pas fiable, pouvant conduire à l'apport de faux polymorphismes dans le gène étudié suite à la génération du consensus. La probabilité de ce phénomène est d'autant plus élevée que la taille des lectures est petite et la région séquencée est répétée. D'une longueur moyenne de 150 bases, ces séquences sont plus susceptibles que les longues lectures de correspondre à la fois à la région d'intérêt et à d'autres régions du génome, et d'avoir été sélectionnée lors de l'extraction faite par Javier BELINCHON MORENO.

3. L'alignement avec les **longues et courtes lectures** à la fois produit un consensus tout aussi complet. Cependant, ce scénario pose également problème dans la mesure où les courtes lectures sont largement majoritaires en couverture par rapport aux longues lectures et prendraient le dessus lors de la création du consensus. Ainsi, la problématique d'apport de polymorphismes d'autres régions du génome par un faux mapping des courtes lectures resterait inchangée. De plus, les longues et courtes lectures ayant des conditions de fiabilité et de couverture différentes, l'idéal serait de pouvoir ajuster les paramètres d'alignement différemment pour chaque type de lecture. Cela permettrait d'avoir davantage de finesse sur les commandes, ce qui n'est cependant pas possible en les regroupant dans un même alignement car les paramètres sont les mêmes pour tous les fichiers sélectionnés.

4. Enfin, les alignements avec les **longues puis courtes** permettent d'obtenir un consensus conforme à l'attendu. Cette démarche d'alignement des courtes lectures sur un consensus issu de longues lectures se nomme un *polishing*. Elle permet de "lisser" les potentielles erreurs de séquençage et d'alignement des longues lectures en palliant le manque de couverture ou le manque de fiabilité du séquençage NAS. Le fait d'utiliser le consensus NAS comme référence pour l'alignement des courtes lectures permet d'être plus spécifique de la nouvelle variété et ainsi de limiter les faux mapping de courtes lectures d'autres régions génomiques extraites d'après Anso77. Ainsi, la suspicion d'apport de faux polymorphismes est fortement diminuée.

Parmi les 4 consensus, on observe que celui obtenu avec le scénario 4 (longues puis courtes lectures) correspond le mieux à nos attentes en termes de couverture, code IUPAC et correspondance témoin positif / référence. Ayant confirmé cette démarche la région Chr10, j'ai retenu le scénario 4 pour la reconstruction des régions d'intérêt pour l'ensemble des variétés.

Ainsi, je réalise d'abord l'alignement uniquement des longues lectures sur la référence Anso77, puis je génère les consensus correspondants à chaque nouvelle variété (consensus NAS). Ensuite, je réalise le polishing avec les courtes lectures en les alignant sur les consensus NAS devenus alors les nouvelles références. Cet alignement permet de générer les consensus finaux pour chaque variété. La ligne conductrice de la démarche est décrite par le schéma simplifié en ***Figure 9*** .

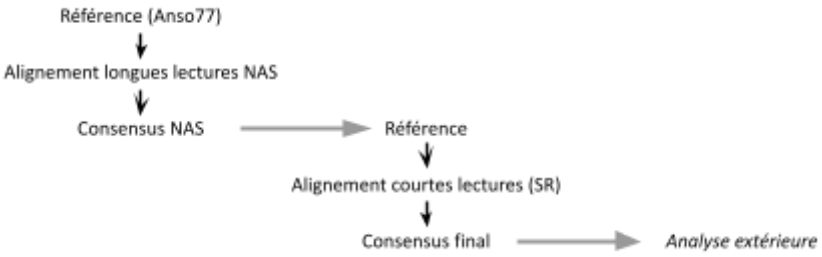


Figure 9 : Schéma de la démarche de reconstruction des consensus

Les profondeurs observées en longues lectures complètes confirment l'hypothèse de départ de similarité de séquences entre les 6 régions d'intérêt et leur environnement, et les séquences des autres régions premièrement étudiées.

La démarche réalisée présente tout de même quelques inconvénients. D'une part, on observe la présence de code IUPAC dégénéré dans les consensus d'Anso77 (témoin positif). Cela peut s'expliquer par le fait que la référence et le séquençage pour obtenir les lectures du témoin positif aient été produits à partir de 2 individus différents. Bien qu'appartenant à la même variété, il est possible d'observer quelques polymorphismes, d'autant plus dans ces régions complexes des gènes NLR. D'autre part, sur l'ensemble du jeu test, on remarque de fortes variations de couverture qui peuvent être dues à la méthode de sélection des lectures basée sur la similarité avec la référence Anso77. Le choix de la combinaison de paramètres adaptés permet par la suite de rectifier ces points problématiques.

III.2 Étalonnage des paramètres

Suite à l'analyse des tests avec les 2 variétés (Anso77 et Moquatre) ainsi que quelques ajustement lors des essais workflows, j'ai défini les paramètres optimums d'alignement et l'analyse de l'arbre de décision m'a permis d'établir les paramètres de construction des consensus.

Paramètres d'alignement :

➢ Qualité des bases

Aucune qualité minimale pour les bases n'était requise dans les paramètres par défaut. J'ai donc placé cette dernière à 30, ce qui correspond à une qualité de 99,9% (en comparaison, une qualité de 10 correspond à 90% et 20 à 99%). La plupart des lectures étant déjà de bonne qualité, cela permet d'éliminer les lectures contenant des erreurs de séquençage sans pour autant trop réduire la couverture.

➢ Taux de mismatchs

Le paramètre "Minimum Mismatch" (MM) définit le pourcentage de mismatchs autorisé sur la longueur de chaque lecture lors de l'alignement. Il est fixé à 30% par défaut. Plus le MM accepté est élevé, plus il y a de risques que la lecture n'appartienne pas à la région d'intérêt mais ait été intégrée car elle s'aligne en partie à la référence, ou bien que son séquençage ait été de mauvaise qualité. Mais ces mismatchs peuvent également correspondre à des polymorphismes (SNP et insertions/délétions), il ne faut donc pas mettre ce paramètre à 0% car l'objet de cette étude est justement de repérer ces polymorphismes.

J'ai remarqué qu'il était mieux que le MM soit différent en fonction de l'alignement longues ou courtes lectures. Concernant les longues lectures, elles peuvent contenir des erreurs de séquençage et sont la base du premier consensus de la nouvelle variété. Il faut ainsi tolérer un nombre élevé de mismatch pouvant correspondre à de vrais polymorphismes. J'ai donc accordé un MM de 15%. Les courtes lectures servent à corriger les potentielles erreurs des longues lectures et pallier le manque de couverture. Ainsi, le MM n'a pas besoin d'être élevé. Pour ce 2ème alignement, avec les courtes lectures, j'ai donc placé le MM à 5%.

➢ Taille maximale des ouvertures

Avec une taille maximale de 50 bases fixée par défaut par Geneious Prime, les insertions de tailles supérieures n'auraient pas été observées lors de l'alignement, ce qui est incompatible avec une recherche de polymorphismes de structure entre les variétés et la référence. J'ai donc fixé la taille maximale pour les ouvertures à 500 bases, permettant d'identifier de grandes insertions/délétions tout en limitant le risque de faux mapping en ouvrant de trop.

Paramètres de consensus :

➢ Couverture

En l'absence de couverture lors de l'alignement sur la séquence de référence, Geneious Prime place un "?" dans le consensus car le manque de lectures ne permet pas de savoir si la variété étudiée conserve les mêmes bases que la référence ou si un polymorphisme est potentiellement présent. Le logiciel indique que le consensus n'est pas décisif à cette position. Cependant, tout comme les lettres indésirables du code IUPAC, le "?" n'est pas reconnu par les autres logiciels et sites en ligne pour la comparaison de séquences. Pour obtenir le consensus final de la nouvelle variété exploitable en dehors de Geneious Prime, il a donc fallu choisir entre remplacer "?" par "N", qui est la lettre universellement reconnue en cas d'absence de donnée pour une position, ou bien conserver la séquence de la référence.

Pour cela, j'ai réalisé un arbre de décision (***Annexe 9***) présentant tous les cas possiblement observables de couverture longues et/ou courtes lectures, les hypothèses de leurs significations et les résultats attendus pour différentes combinaisons de paramètres. Toutes les combinaisons de paramètres possèdent des points négatifs, certains plus contraignants que d'autres. J'ai choisi le scénario amenant le moins d'erreurs et étant optimal pour la majorité des cas. Le consensus final a donc été reconstruit en mettant la référence en cas d'absence de couverture longues lectures lors du premier alignement. Ceci permet d'optimiser l'alignement des courtes lectures sur le consensus NAS, qui pourrait être limité en présence de N dans la référence. Suite à cela, des N seront placés en cas d'absence de courtes lectures. Je limite ainsi au maximum l'insertion de séquence d'Anso77 dans le consensus final des variétés.

➢ Seuil de consensus

Le seuil de consensus programmé à 65% par défaut , permet de conserver la base majoritaire à 65% pour chaque position dans le consensus. Ce pourcentage entraîne une haute fiabilité du consensus mais pose problème dans le cas d'une couverture peu élevée où on peut observer qu'une base domine en majorité mais pas suffisamment pour atteindre 65%. Le consensus affiche alors le code IUPAC dégénéré. Après plusieurs tests, j'ai d'abord conclu qu'il est préférable de mettre ce paramètre à 50% afin d'obtenir un consensus à partir de la base majoritaire à 50% pour chaque position. Cependant, ce seuil est toujours problématique en cas d'hésitation entre 3 bases où aucune n'atteint 50% (par exemple à 25%, 30% et 45%). Après concertation avec mes responsables, il a été décidé qu'il valait mieux fixer le seuil de consensus à partir de la base majoritaire, peu importe le pourcentage. Cette option est appelée "0% majority" sur Geneious Prime.

➢ Extraction du consensus de la région d'intérêt

Lors de la phase de tests manuels, pour générer les consensus (NAS et final) de la région d'intérêt pour la nouvelle variété, j'ai réalisé manuellement l'étape d'extraction de la partie de l'alignement correspondant à la région d'intérêt à la base près. J'ai ensuite pu automatiser cette étape dans le workflow en sélectionnant l'option "Trim to reference region". En éliminant ces fragments de séquences non alignées sur la région d'intérêt, la comparaison ultérieure du consensus final de chaque variété sera facilitée.

III.3 Résultats des tests workflow

J'ai pris en main la construction du workflow avec des commandes simples (***Figure 10***). Ce dernier permet de réaliser les actions suivantes : "Parmi tous les documents sélectionnés en entrée (1), sélectionner uniquement ceux dont le nom contient '*reads_NAS*' (2) puis les aligner d'après la référence et les paramètres choisis dans la commande (3)."

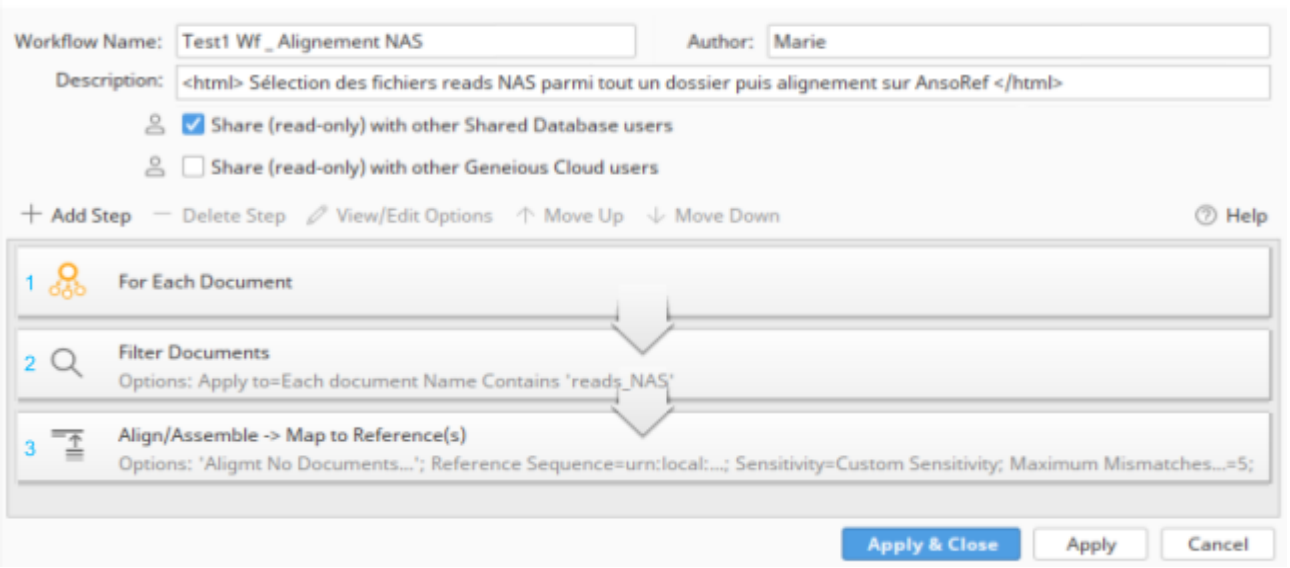


Figure 10 : Premier workflow valide créé

J'ai ensuite complexifié les workflows et effectué de nombreux tests avant d'arriver au workflow final de reconstruction complète des régions génomiques d'intérêt. Le workflow final suit la démarche décrite dans le paragraphe *III.1 Démarche de reconstruction*. Cette ligne conductrice du workflow est relativement simple. Cependant, automatiser l'enchaînement de ces étapes nécessite de nombreuses commandes intermédiaires, telles que renommer des fichiers pour qu'ils soient bien reconnus, enregistrer la progression, lier une étape à une précédente, etc. Dans un premier temps, j'ai créé et validé 2 workflow indépendants lancés successivement (***Figures 11.a et 11.b***). Leur fusion, qui impose l'ajout d'un nombre important d'étapes n'ayant pas abouti comme espérée, j'ai contacté le service support de Geneious Prime. Après discussion, j'ai testé leurs alternatives sans succès. Dans le temps imparti du stage, il a été plus simple de continuer à lancer les 2 workflows l'un à la suite de l'autre. Voici donc un résumé des commandes du workflow final :

- Étape 1
    1. Alignements des longues lectures NAS de chaque variété sur la référence Anso77 (+ sauvegarde)
    2. Consensus de cet alignement longues lectures NAS pour chaque variété (+ sauvegarde)
    3. Raccourcissement du nom du consensus NAS pour un plus court et changement du nom des courtes lectures pour qu'il inclut le nom du consensus NAS en préfixe (+ sauvegarde)
- Étape 2 (polishing)
    4. Alignement des courtes lectures de chaque variété sur le consensus NAS correspondant à la même variété en fonction du nom (préfixe) des fichiers (+ sauvegarde)
    5. Consensus de cet alignement courtes lectures pour chaque variété (+ sauvegarde)

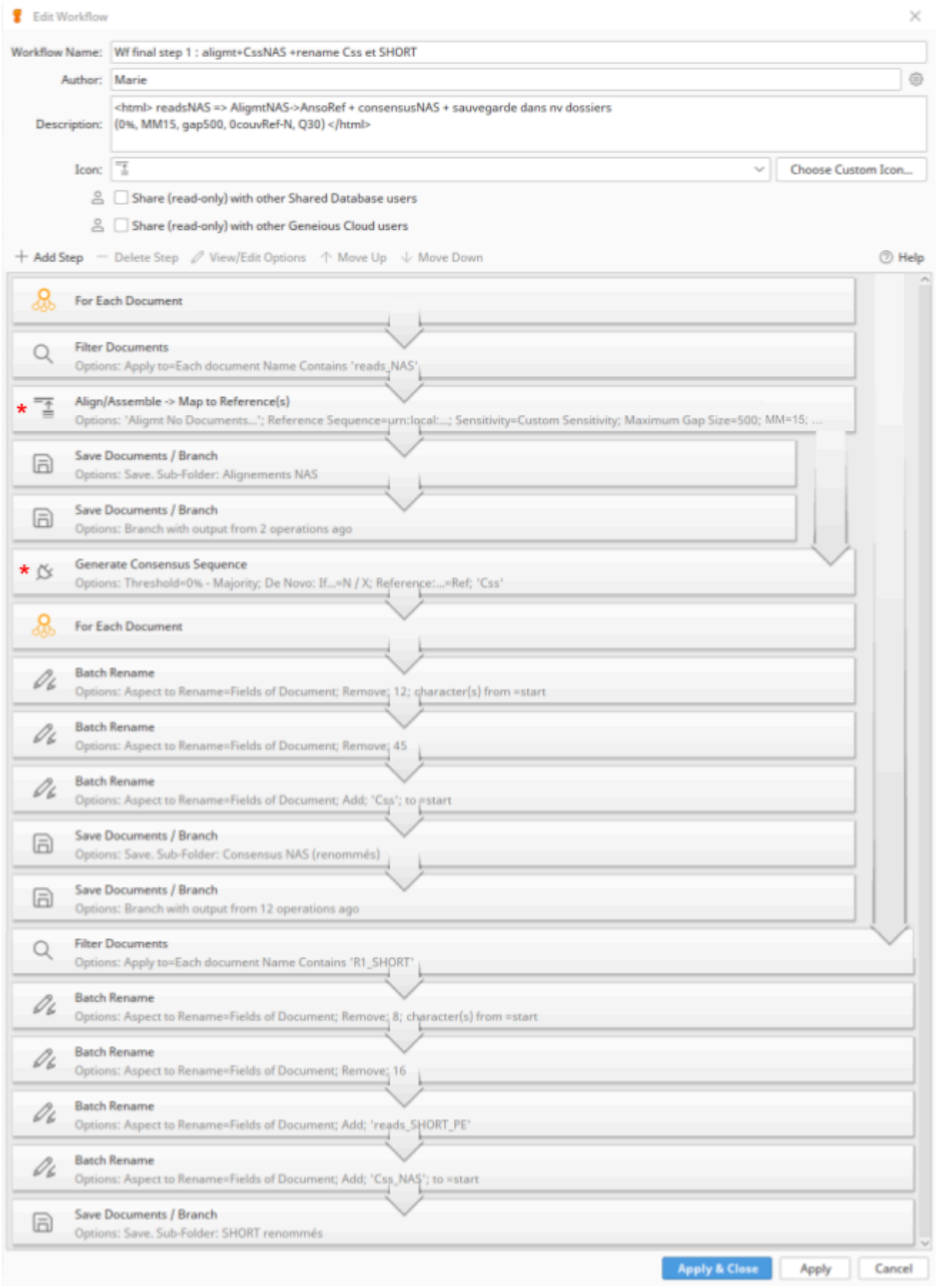


Figure 11.a : Première étape du workflow final avec Geneious Prime

Les étape dotées d'un * correspondent à la ligne conductrice, hors commandes intermédiaires.

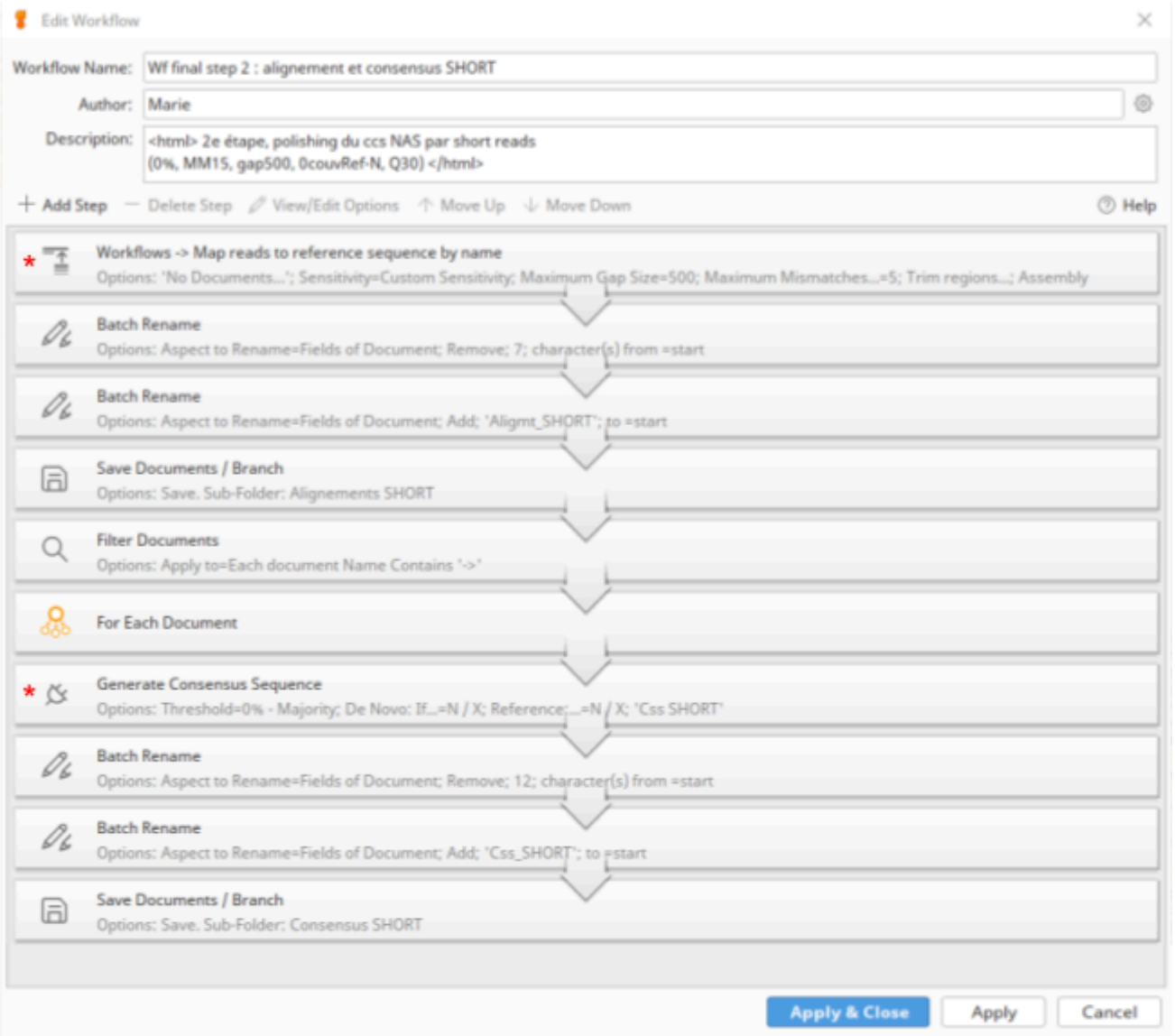


Figure 11.b : Deuxième étape du workflow final avec Geneious Prime

Les étape dotées d'un * correspondent à la ligne conductrice, hors commandes intermédiaires.

Suite à la construction du workflow, j'ai pu facilement lancer l'analyse sur 7 nouvelles variétés. Ces tests des workflows ont permis de visualiser des cas envisagés mais absents d'Anso77 et Moquatre. En effet, ces 2 variétés sélectionnées pour les tests d'étalonnage n'étaient pas représentatives de tous les cas problématiques de couverture possibles. La diversité du jeu test complet m'a donc permis de confirmer les paramètres d'alignement et de consensus pour la couverture, choisis suite à la construction de l'arbre de décision.

J'ai donc reconstruit les consensus finaux des 9 premières variétés (***Tableau 4***). Ces derniers sont variés en termes de polymorphismes structuraux et nucléotidiques, visibles d'après la taille des consensus.

Tableau 4 : Caractéristiques des consensus finaux des variétés du jeu test pour les 6 régions d'intérêt, obtenus avec les paramètres adaptés du workflow

Régions	Taille référence	Caractéristiques	Variétés								
			ANSO77	MOQUATRE	EARLS88632	K2005	PI116479	HSD753	SHENDI	FQUS	CUM413
Chr6a Gène 1	42 315	Taille (en bases)	42 327	42 187	42 186	42 308	42 293	42 520	42 333	42 195	42 185
		Code IUPAC	1W, 1S, 1K	0	0	1R, 86N, 1M	1W	17N, 2Y, 1W, 2S	0	0	2Y, 1R
Chr6b Gène 2	43 170	Taille (en bases)	43 173	43 103	43 108	42 829	43 170	42 939	42 924	43 179	43 176
		Code IUPAC	13N	0	0	39N, 1K, 1M, 9R, 4S, 1W, 9Y	0	12N	0	1N 1Y	1N 1R
Chr9a Gène 3	41 233	Taille (en bases)	41 230	40 994	41 154	41 160	40 615	41 278	41 280	41 163	41 210
		Code IUPAC	2 212 N	2 397 N, 3Y, 2W	2 296 N, 4W	2 157 N, 2W	2 177 N, 1R	2 259 N	2 231 N, 1W	2 209 N	2 195 N, 1Y
Chr9b Gène 4	44 768	Taille (en bases)	45 101	44 775	44 810	44 643	44 784	45 678	44 816	44 860	44 779
		Code IUPAC	9 628 N	9 277 N	9 308 N	9 258 N, 1R	9 286 N, 1Y, 1K, 1M, 1R,	9 437 N	9 353 N	9 326 N, 2W, 1R	9 267 N, 1W
Chr10 Gène 5	40 448	Taille (en bases)	40 477	40 406	39 357	39 735	40 215	39 643	40 401	40 469	39 650
		Code IUPAC	1 602N, 1K, 1W	8 368N, 5Y, 2W, 1S, 4R, 2M, 1K	6 357N, 1Y, 5W, 2R, 1M	4 008N, 6Y, 3W, 9R, 1M, 4K, 1H	15 289N, 5Y, 6W, 1S, 5R, 1M, 3K	5 048N, 8Y, 9W, 8R, 1H	1 575N, 11W, 1R, 1M, 1K	1 436N, 4Y, 7W, 1R	2 070N, 3Y, 1S, 6R, 1M, 1K
Chr12 Gène 6	40 809	Taille (en bases)	40 810	40 812	40 823	41 161	40 789	40 658	NA	41 332	41 326
		Code IUPAC	1 379N	1 349 N	1 497N	1 370N, 1W, 1K	1 389N	1 647 N, 4W, 1S	NA	1 562N	1 429N

La similarité des tailles de consensus d'Anso77 et de la référence valide le workflow de construction des consensus des régions d'intérêt. La variation de taille pour les autres variétés correspond alors à des polymorphismes de structure (insertions / délétions).

Concernant la variété Shendi en région Chr12, l'alignement NAS a produit 2 consensus NAS, bloquant ainsi la suite du workflow avec l'alignement courtes lectures. L'hypothèse d'une forte variation structurale (insertions ou délétion) pourra être étudiée en perspective.

On remarque un très grand nombre de N dans les consensus finaux de toutes les variétés pour les régions Chr9a, Chr9b, Chr10 et Chr12. Dans les séquences des consensus, les N co-localisent aux extrémités. Après vérification des alignements, on observe bien une absence totale de couverture à ces positions.

Ce phénomène s'explique par l'utilisation de 2 fichiers différents de régions d'intérêt (gène ±20kb) pour l'extraction des lectures et comme référence. Une erreur ayant été remarquée sur le 1er fichier des régions d'intérêt (extraction des lectures sur moins de 40kb) , un 2ème fichier a été produit, conduisant à une référence plus longue que ce que les lectures ne peuvent couvrir. Dans un premier temps, ces N ne sont pas problématiques pour l'analyse finale car ils ne se situent pas à proximité des gènes d'intérêt, attendus au milieu de la séquence à partir de 20kb.

Concernant le reste du code IUPAC, on remarque qu'Anso77 en possède pas ou très peu en fonction des régions. Dans le cas des autres variétés, en cas de présence de bases dégénérées, on peut émettre l'hypothèse d'une hétérozygotie résiduelle ou de duplication. Bien que le melon soit homozygote, il est possible que des traces de deux allèles pour un même gènes, conduisant à cette incertitude de consensus. Quant à la duplication, si une variété possède deux copies d'un gène unique chez Anso77, les 2 copies de ce gène s'alignent aux mêmes positions de la référence.

III.4 Annotation des séquences consensus finales

La reconstruction des consensus finaux des régions d'intérêt a pour objectif d'étudier le polymorphisme des 6 gènes de résistance, Je me suis donc penchée sur l'annotation des régions d'intérêt à l'aide de la plateforme en ligne BLAST (Basic Local Alignment Search Tool) [BlastX] pour localiser des motifs caractéristiques des gènes de résistance.

Les 10 résultats de BLASTx correspondant le mieux à la séquence Chr6a:19000-25000 de Moquatre sont présentés dans la ***Figure 12*** . D'après la base de données du NCBI, Anso77 et Moquatre ont une forte probabilité de contenir, en région Chr6a:19000-25000, un domaine protéique TIR caractéristique des gènes NLR. Une annotation expertisée serait nécessaire pour conclure si ces 2 variétés possèdent un gène de résistance d'intérêt ou un pseudogène (motif TIR isolé).

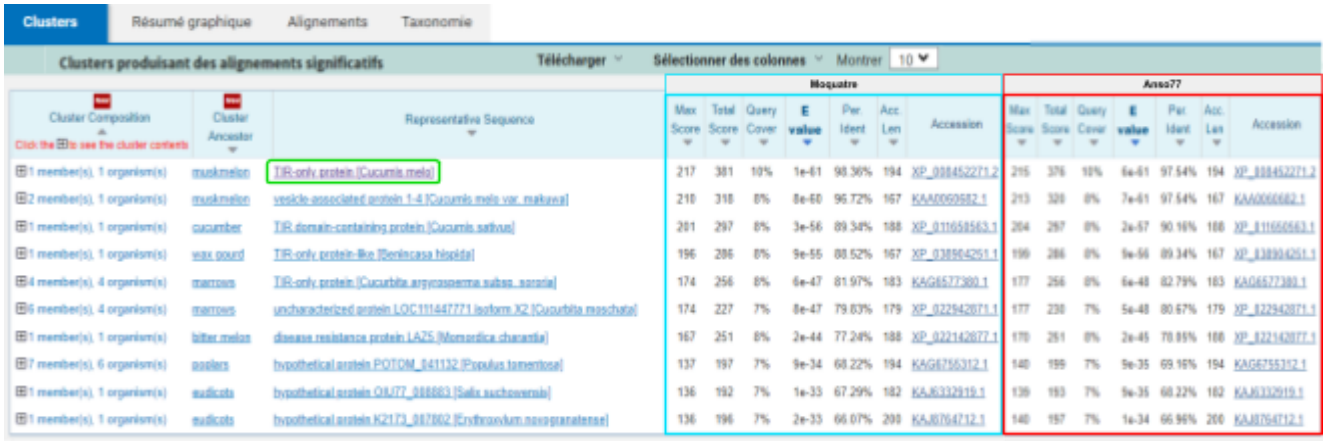


Figure 12 : Résultats BLASTx : correspondance des séquences d'Anso77 et Moquatre pour la région Chr6a (gène 1) et des protéines connues dans les bases de données

IV. Conclusion et perspectives

IV.1 Conclusion de la problématique de stage

Pour conclure, l'objectif de mon stage a été atteint. Chargée d'établir la démarche de reconstruction de régions d'intérêt pour 144 variétés, j'ai réussi à construire un workflow permettant de générer automatiquement les alignements et consensus des variétés à partir d'une sélection de fichiers longues et courtes lectures. Ce travail s'est déroulé en plusieurs étapes, rythmées de réflexions et prises de décision.

Les 2-3 premières semaines ont été un temps d'appropriation de Geneious Prime pour comprendre son fonctionnement et repérer les avantages et inconvénients vis-à-vis de mon objectif d'étude. Durant cette phase, j'ai utilisé 2 variétés parmi les 144 en tant que témoin positif et test. L'obtention des premiers consensus et l'analyse des polymorphismes, m'ont permis de définir les paramètres optimaux pour obtenir des séquences consensus fiables et représentatives des nouvelles variétés. J'ai ainsi établi la ligne conductrice de la démarche de reconstruction complète des régions d'intérêt.

Ensuite, la mise en place du workflow, de l'élaboration de la suite de commandes aux tests pour l'ajustement des paramètres, m'a occupée 1 mois. La diversité des variétés testées a permis de valider les paramètres préalablement définis. En conclusion de cette étape, le workflow créé est fonctionnel et a permis de reconstruire les 6 régions des 9 premières variétés.

Enfin, l'annotation et l'étude des polymorphismes des séquences d'intérêt est l'aboutissement de mon travail. J'ai commencé l'annotation avec l'outil BlastX en comparant les séquences consensus finales d'Anso77 et Moquatre (région Chr6a) à des bases de données publiques en génomique, et confirmer la présence d'un motif TIR.

IV.2 Perspectives de l'étude

Mon stage étant prévu pour 4 mois, il me reste 2 mois pour continuer moi-même mon travail pendant un certain temps. Plusieurs objectifs restent à atteindre avant l'intégration de cette étude à la thèse de Javier BELINCHON MORENO.

Tout d'abord, je vais trouver une issue pour le workflow lorsque plus d'un consensus NAS est produit. En parallèle, l'extraction des séquences longues et courtes lectures avec le 2ème fichier de régions d'intérêt sera relancée pour obtenir des consensus finaux avec un minimum de N. L'efficacité du workflow me permettra de réaliser cette étape rapidement. Puis je poursuivrai l'annotation et la comparaison des séquences pour les variétés du jeu test. Je réaliserai les BLAST de toutes les séquences consensus finales des 9 variétés pour la recherche de motifs protéiques caractéristiques des gènes de résistance NLR. J'utiliserai également une plateforme bioinformatique DNA subway permettant l'annotation de génome. Lors des premières semaines après mon arrivée, j'ai eu l'occasion, par ma maître de stage, de visualiser quelques annotations à l'aide de cet outil mais n'ai pas pu tester par la suite cette méthode sur les variétés d'intérêt dans le temps imparti de mon stage. Il se peut que j'avance dans cette voie d'annotation et de comparaison d'ici l'oral de stage et que j'ai de nouveaux résultats à présenter.

Ensuite, j'étendrai la reconstruction de régions d'intérêt aux 135 autres variétés. Pour cela, il me suffira de lancer le workflow déjà créé, les paramètres ayant été réfléchis et définis au cours de ces 2 premiers mois de stage. Pour cela, Javier BELINCHON MORENO me transmettra les fichiers de longues et courtes lectures pour les 135 autres variétés.

Pour finir, je poursuivrai mon travail jusqu'à l'analyse des consensus finaux de ces 135 variétés, comme effectué avec le jeu test. Des annotations et comparaisons de séquences seront alors nécessaires.

Malgré la simplicité de ces perspectives de travail, je ne peux pas en prédire précisément le déroulé. En effet, au cours de ces 2 premiers mois de stage, de nouvelles problématiques et questionnements ressortaient régulièrement. Il est donc probable que je rencontre de nouveaux obstacles avec les 135 dernières variétés ou avec l'annotation des régions. Par exemple, il est possible que certaines variétés n'aient pas suffisamment de couverture pour réaliser un consensus acceptable, même avec un grand nombre de N. Il faudra donc envisager soit de revoir l'extraction des lectures, réaliser une PCR ou autre.

IV.3 Retour d'expérience du stage

Ce stage a été une expérience très enrichissante, tant du point de vue professionnel que social.

Pour cette première expérience professionnelle, l'intégration sociale dans l'entreprise ressort en premier lieu. J'ai eu la chance d'intégrer une équipe bienveillante et investie dans le suivi de mon travail. En dehors du cadre professionnel de mon stage, ils m'ont naturellement fait sentir intégrée lors des pauses, devenues des moments de partage.

Ce stage m'a beaucoup apporté sur le plan professionnel. D'une part, j'ai acquis de nombreuses connaissances dans le domaine de la génomique et ai eu l'occasion de m'immerger dans le monde de la recherche. D'autres part, j'ai pu affiner ma méthode pour la structure d'un rapport scientifique grâce à la lecture de nombreuses publications scientifiques et par l'aide précieuse de mes responsables de stage pour l'écriture de ce rapport-même.

Un des points que j'ai particulièrement apprécié au cours de ce stage a été l'omniprésence de l'anglais dans mon travail. Je ne disposais pratiquement que de documentations en anglais, que ce soient les thèses et articles lues en début de stage pour comprendre le contexte et pour l'écriture finale de mon rapport, ou encore le manuel d'utilisation du logiciel Geneious Prime et les échanges avec le service support. J'ai également assisté à des conférences et animations en anglais.

Les difficultés rencontrées au cours de ce stage ont été dans mon travail lui-même et les missions confiées. Tout d'abord, les premières semaines de stage ont été éprouvantes à travers la charge intellectuelle. Il m'a fallu un peu de temps pour me plonger dans le domaine de la génomique végétale, peu étudié à mon niveau de formation, et suivre pleinement les réunions d'équipe malgré la différence de familiarité au contexte entre mes responsables et moi.
Ensuite, il a parfois été difficile à l'échelle de l'équipe de trouver des solutions à des problèmes rencontrés au cours du travail. En effet, l'absence de référence ou de connaissance pour la progression du travail est une caractéristique même de la recherche. Je reconnais tout de même un certain intérêt pour cet univers de travail.

Enfin, la vulgarisation pour le rapport de stage n'a pas été si simple car il m'a fallu expliquer en détails mais clairement des notions complexes sur lesquelles je travaille depuis peu. Ayant commencé l'écriture dès le premier mois, cela m'a permis de prendre du recul sur toutes ces nouvelles connaissances et ainsi améliorer ma compréhension en organisant mes idées.

En dehors du cadre de mes missions de stage, j'ai eu la chance de vivre quelques expériences enrichissantes.

Au cours du premier mois de stage, j'ai pu visiter les laboratoires, me permettant entre autres de visualiser les technologies de séquençage ayant permis l'obtention des données génomiques utilisées au cours de mon stage. J'ai également eu l'occasion de réaliser le dépôt d'une banque d'ADN sur une flowcell pour un séquençage NAS dans le cadre d'un projet parallèle dans l'unité.

Le 25 avril, j'ai participé au DNAday au Génopole (à Évry). Lors de cet événement, de nombreuses présentations ont été animées telles que des rappels de l'historique des recherches en génomique et de la création du Génopole mais aussi des projets d'entreprises comme Medicen et DNA Script. J'ai assisté à une visite d'Illumina, la société de séquençage courtes lectures, où il nous a été présenté les différents séquenceurs et technologies fabriquées par cette société.

Ces conférences et visites étaient entièrement en anglais, ce qui m'a permis de me plonger dans cette langue à l'oral également.

En conclusion personnelle de stage, je ne pouvais pas espérer de meilleures conditions pour cette première expérience professionnelle. J'en garderai un très bon souvenir et des compétences utiles pour la suite de mon chemin dans le monde professionnel.

Abréviations et définitions

ADN : Acide DésoxyriboNucléique

↪ macromolécule biologique contenant l'information génétique (génome) et présente dans presque toutes les cellules et certains virus.

BAP : Biologie et Amélioration des Plantes

↪ département de l'INRAE dont dépend l'unité EPGV

BLAST : Basic Local Alignment Search Tool

↪ méthode de recherche heuristique utilisée en bioinformatique pour trouver les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés, et de réaliser un alignement de ces régions homologues. La ou les séquences introduites par l'utilisateur sont comparées avec des séquences répertoriées dans des bases de données ayant des similitudes.

CC : Coiled Coil (en français, superhélice)

↪ domaine protéique

CEA : Commissariat à l'Énergie Atomique et aux Énergies Alternatives

↪ organisme de recherche scientifique français dans les domaines de l'énergie, la défense, la santé, les technologies de l'information et de la communication, les sciences de la matière et de la vie.

CNRGH : Centre National de Recherche en Génomique Humaine

EPGV : Étude du Polymorphisme des Génomes Végétaux

↪ unité d'INRAE

GAFL : Génétique et Amélioration des Fruits et Légumes

↪ unité d'INRAE dont dépend Javier BELINCHON-MORENO pour sa thèse, localisée à Avignon

GWAS : Genome-Wide Association Study (Étude d'association pangénomique, en français)

↪ analyse de variations génétiques en corrélation avec des traits phénotypiques

Illumina ⟶ société américaine de séquençage courtes lectures en single end ou paired-end

INRAE : Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

IUPAC : International Union of Pure and Applied Chemistry

↪ organisation non gouvernementale (membre du Conseil international pour la science) qui s'intéresse aux progrès en chimie, chimie physique, biochimie… et dont l'autorité est reconnue pour le développement de règles à adopter pour la nomenclature, les symboles et la terminologie des éléments chimiques et de leurs dérivés.

kb, Mb, Gb : kilo bases (millier), méga bases (million), giga bases (milliard)

↪ unités de mesure en nombre de bases pour la longueur des séquences

MGI : filiale du BGI (Beijing Genomic Institute)

↪ société chinoise de séquençage courtes lectures en single-end ou paired-end

NAS : Nanopore Adaptive Sampling (en francais, Séquençage Adaptatif Nanopore)

↪ technique de séquençage longues lectures par nanopore pour des régions cibles

NBS LRR : Nucleotide-Binding Site Leucine-Rich-Repeat (abrégé en **NLR**)

↪ famille de gènes dans laquelle on retrouve la majorité des gènes de résistance

NCBI : National Center for Biotechnology Information

↪ institut national américain pour l'information biologique moléculaire et banque de données en libre accès

ONT : Oxford Nanopore Technologies

↪ société de séquençage longues lectures par nanopore pour tout génome (sans sélection de région)

SNP : Single Nucleotide Polymorphism

↪ variation génétique d'une seule ou quelques nucléotide(s)

SR : Short Read

↪ courte lecture d'ADN

TIR : Toll-Interleukin Receptor

↪ domaine protéique

Liste des tableaux et figures

Tableaux

Figures

Bibliographie

Pour obtenir des informations complémentaires sur les documents (auteur, date, publication…), je me suis aidée du site suivant : Semantic Scholar | AI-Powered Research Tool.
NB : Certains des documents ci-dessous ne sont pas accessibles en ligne gratuitement mais j'ai pu y accéder grâce à mes identifiants Inrae ou bien je disposais de la version papier.
* Documents lus au cours des premiers jours de stage pour comprendre le contexte génomique de l'étude et du stage.

[1] **Liste des départements INRAE**
Page web, *inra.fr*
https://www.static.inrae.fr/departements

[2] Bernot A.
L'analyse des génomes
Livre, *Nathan Université*, ISBN 2-09-190794-4 (1999)
https://www.decitre.fr/livres/l-analyse-des-genomes-9782091907949.html

[3] **Domaine protéique**
Page web, *Wikipedia.org*
https://fr.wikipedia.org/wiki/Domaine_prot%C3%A9ique

[4] Bresson A.
Caractérisation de R1 et Rus : deux loci de résistance à la rouille foliaire sur le chromosome 19 du peuplier.
Thèse, *Biotechnologies* (2011). Université d'Évry-Val-d'Essonne. Français.
NNT : .tel-02806850

[5] **Immunité des plantes: résistance qualitative et quantitative**
Page web, *ichi.pro*
https://ichi.pro/fr/immunite-des-plantes-resistance-qualitative-et-quantitative-109355967234554

[6] Chen N.
Génomique comparative d'un cluster de gènes de résistance chez le haricot commun et le soja (partie I.B)
Thèse, … (2010)
https://theses.fr/2010PA112371

[7] **Le melon, un délicieux fruit de saison**
Page web, *Futura*
https://www.futura-sciences.com/maison/dossiers/jardinage-melon-delicieux-fruit-saison-2186/page/3/

[8] Mo C, Wang H, Wei M, Zeng Q, Zhang X, Fei Z, Zhang Y, Kong Q
Complete genome assembly provides a high-quality skeleton for pan-NLRome construction in melon *
Article, *The plant Journal* (2024)
doi: 10.1111/tpj.16705

[9] **Le melon : tout savoir pour réussir sa culture**
Page web, *jardiner-malin.fr*
https://www.jardiner-malin.fr/fiche/culture-melon.html#:~:text=On%20trouve%20des%20traces%20de%20culture%20du%20melon,compte%20pr%C3%A8s%20de%201000%20vari%C3%A9t%C3%A9s%20de%20melons%20diff%C3%A9rents.

[10] Longlan XU, Yuhua He, Lingli Tang, Yongyang Xu, Guangwei Zhao
Genetics, Genomics, and Breeding in Melon
Article, *Agronomy* (18 November 2022)
https://www.mdpi.com/2073-4395/12/11/2891

[11] **Melon : le problème croissant des virus**
Page web, *reussir.fr*
https://www.reussir.fr/fruits-legumes/melon-le-probleme-croissant-des-virus#:~:text=En%20France%2C%20les%20principaux%20virus%20posant%20probl%C3%A8me%20en,de%20plantes%20connus%2C%20et%20le%20CMV%20%28genre%20cucumovirus%29.

[12] **Melon - Fiches maladies et ravageurs**
Page web, *inra.fr*
https://ephytia.inra.fr/fr/C/7631/Melon-Fiches-maladies-et-ravageurs

[13] Pitrat M.
Disease Resistance in Melon and Its Modification by Molecular Breeding Techniques
Chapitre du livre de Ezura H, Ariizumi T, Garcia-Mas J, Rose J (eds)
Functional Genomics and Biotechnology in Solanaceae and Cucurbitaceae Crops.
Biotechnology in Agriculture and Forestry, vol 70. Springer, Berlin, Heidelberg (2016)
https://doi.org/10.1007/978-3-662-48535-4_11

[14] Université de Batna 02, Faculté des Sciences de la Nature et de la vie, Département d'Ecologie et Environnement
La variabilité de la taille des génomes chez les plantes
Cours de cytogénétique et polyploïdie (2024)
https://staff.univ-batna2.dz/sites/default/files/ayadi_malik/files/cours-chapitre_04-la_variabilite_de_la_taille_des_genomes_chez_les_plantes.pdf

[15] Belinchon-Moreno J, Berard A, Canaguier A, Chovelon V, Cruaud C, Engelen S, Feriche-Linares R, Le-Clainche I, Marande W, Rittener-Ruff V, Lagnel J, Hinsinger D, Boissot N, Faivre Rampant P.
Nanopore adaptive sampling to identify the NLR-gene family in melon (Cucumis melo L.) *
Article, *BioRxiv* (2023)
https://doi.org/10.1101/2023.12.20.572599

[16] Thomas S.
Pressions de sélection exercées par les résistances génétiques du melon sur les populations d'Aphis Gossypii.
Thèse, Université d'Avignon et des Pays de Vaucluse. ffNNT : ff. fftel-02811117 (2011)
https://hal.inrae.fr/tel-02811117/file/45159_20111117100128625_1.pdf

[17] Toda N, Rustenholz C, Baud A, Le Paslier MC, Amselem J, Merdinoglu D, Faivre-Rampant P.
NLGenomeSweeper: A Tool for Genome-Wide NBS-LRR Resistance Gene Identification
Article, *Genes* (Basel). Mar 20;11(3):333 (2020)
doi: 10.3390/genes11030333. PMID: 32245073; PMCID: PMC7141099.

[18] **Production mondiale de melons par pays**
Page web, *AtlasBig.com*
https://www.atlasbig.com/fr-fr/pays-par-production-de-melon

[19] Montel F.
Séquençage de l'ADN par nanopores, Résultats et perspectives
Article, *Médecine/sciences* (16 février 2018)
https://doi.org/10.1051/medsci/20183402014

[20] Loose M, Malla S, Stout M.
Real-time selective sequencing using nanopore technology
Article, *Nature Methods*, 13(9), Art. 9. (2016)
https://doi.org/10.1038/nmeth.3930

[21] **Geneious Prime 2022.1 User Manual** *
Manuel d'utilisation, … (15 mars 2022)
https://assets.geneious.com/documentation/geneious/GeneiousPrimeManual.pdf

[22] **NCBI : National Center for Biotechnology Information**
Base de données, *nih.gov*
https://www.ncbi.nlm.nih.gov/

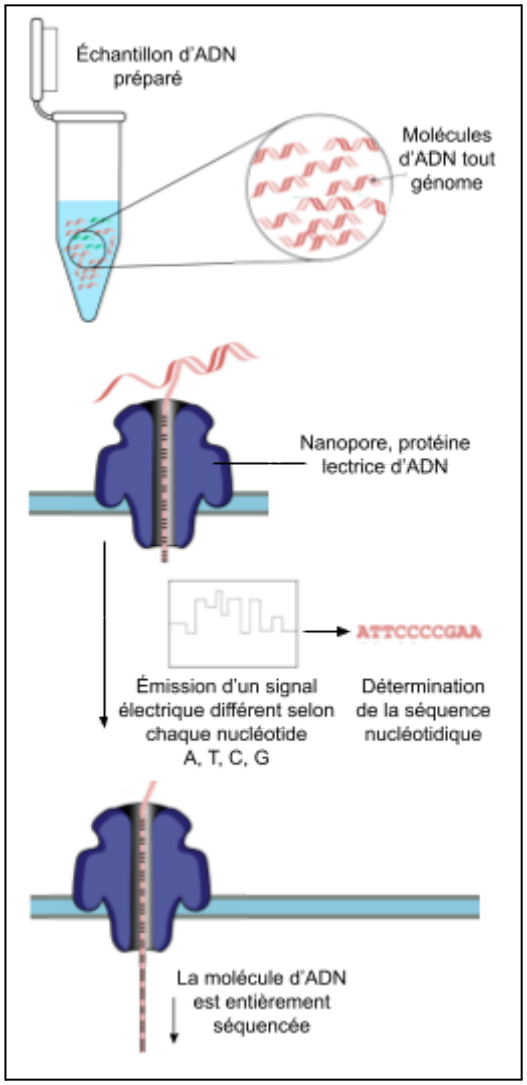
[23] **UniProt**
Base de données, *uniprot.org*
https://www.uniprot.org/

[24] **BLAST: Basic Local Alignment Search Tool**
nih.gov
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&LINK

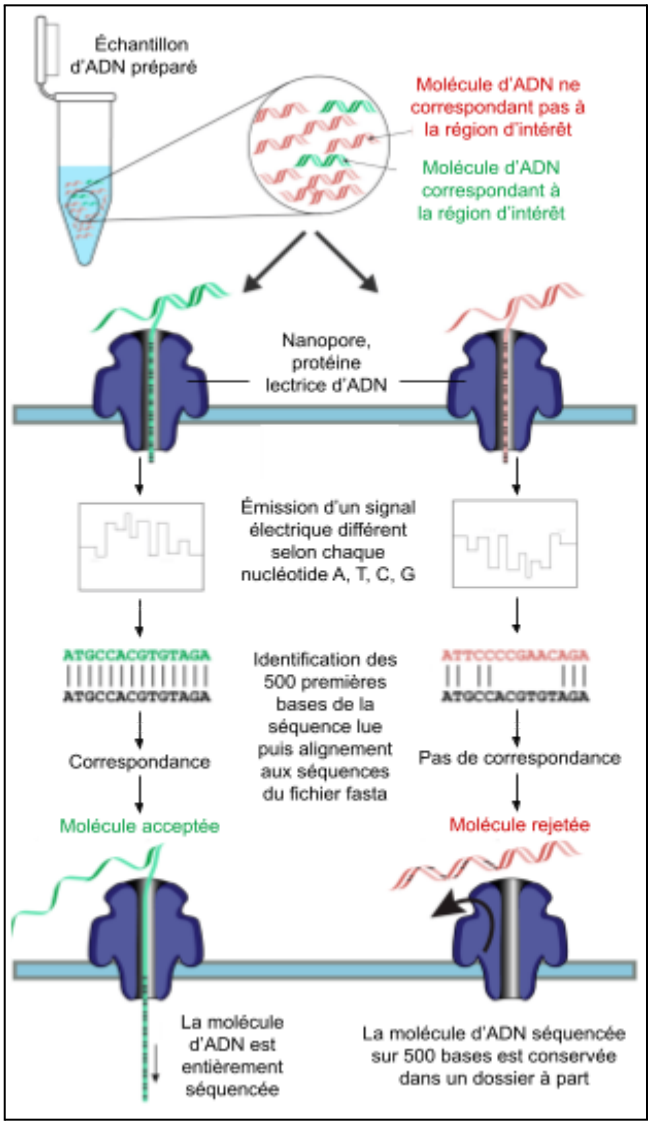
Annexes

Annexe 1 : Groupes botaniques représentés parmi les 144 variétés de melons

- Acidulus
- Adana
- Agrestis
- Ameri
- Cantalupensis
- Chandalak
- Chate
- Chinensis
- Chito
- Conomon
- Dudaim
- Flexuosus
- Inodorus
- Makuwa
- Momordica
- Reticulatus
- Tibish et Seinat
- (Non connu)

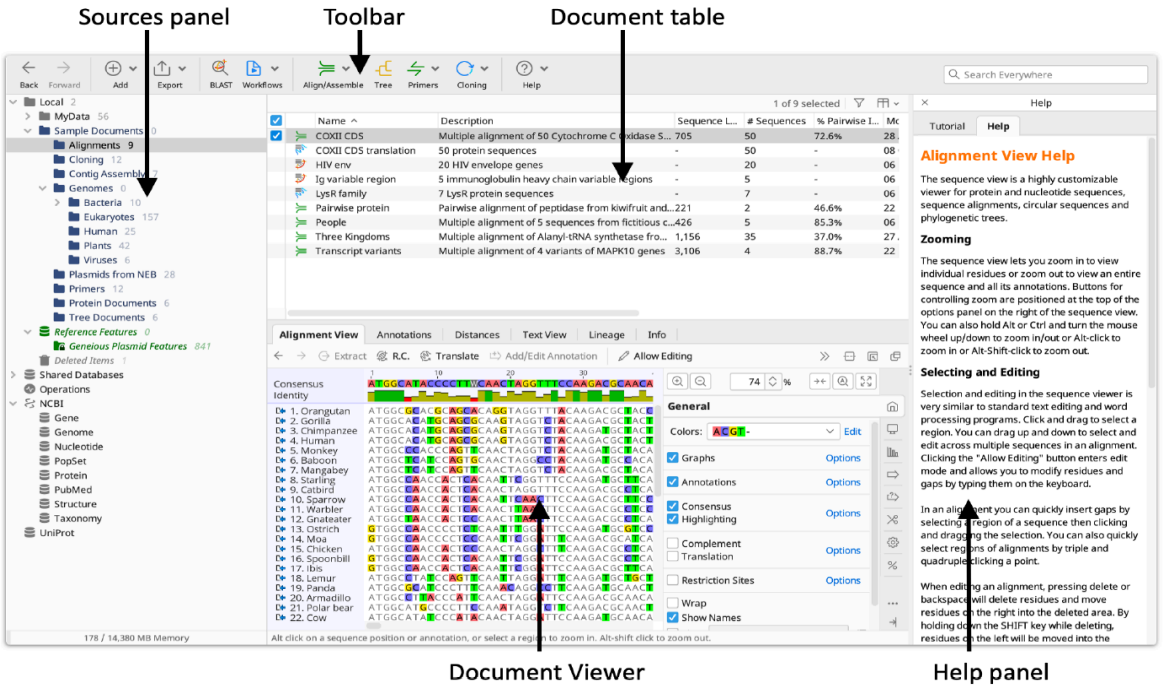


Annexe 2 : Schéma de la méthode de séquençage nanopore standard



Annexe 3 : Schéma de la méthode de séquençage nanopore adaptatif (NAS)

Schémas expliquant la différence entre le séquençage nanopore ONT standard et adaptatif (NAS) pour l'obtention des longues lectures.



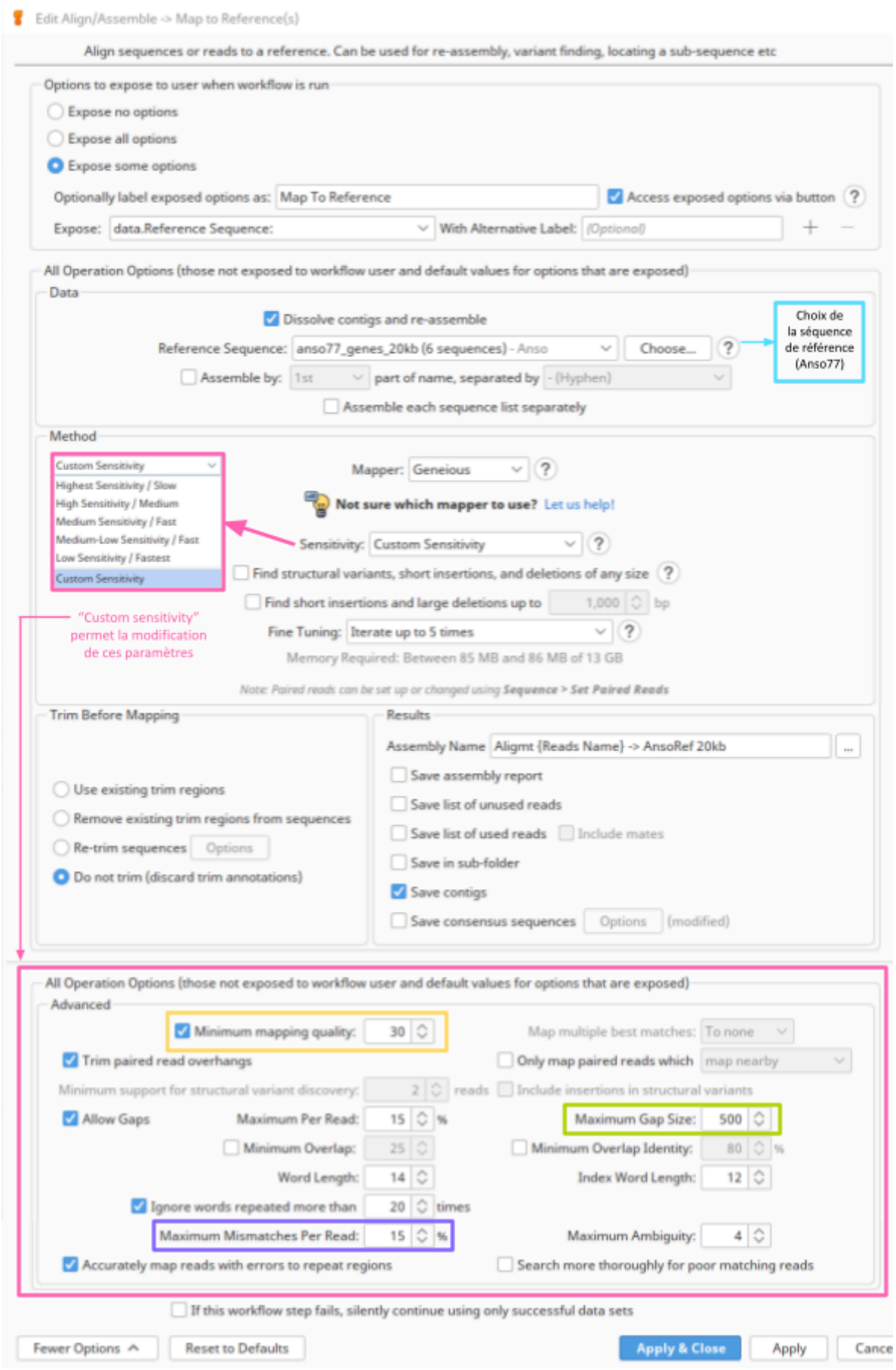
Annexe 4 : Organisation de la fenêtre sur Geneious Prime, indiquant les options disponibles et fichiers (de départ et créés grâce au logiciel)

Annexe 5 : Manuel d'utilisation du logiciel Geneious Prime
https://assets.geneious.com/documentation/geneious/GeneiousPrimeManual.pdf

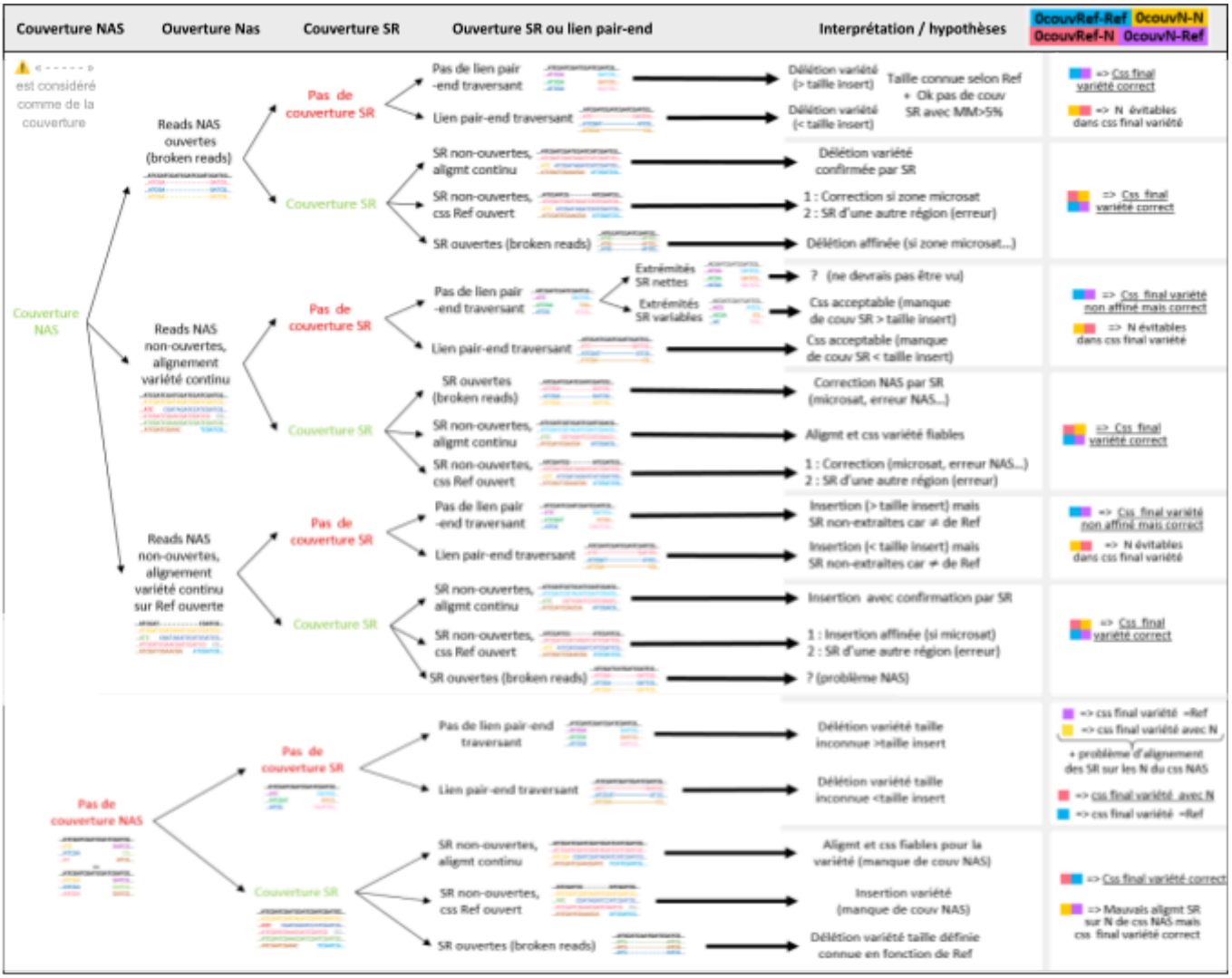
Annexe 6 : Vidéos tutoriel d'utilisation du logiciel Geneious Prime
https://www.geneious.com/academy/

Annexe 7 : Tableau de légende du code IUPAC

Code nucléotide IUPAC	Signification de la base	Code couleur sur Geneious Prime	Conclusion après obtention de la lettre
A	Adénine	A	Accepté
C	Cytosine	C	
G	Guanine	G	
T (ou U)	Thymine (ou Uracil)	T (ou U)	
. ou -	Trou	. ou -	
N	Aucune base ou les 4 à la fois	N	
Y	C ou T	Y	Indésirable
S	G ou C	S	
W	A ou T	W	
K	G ou T	K	
M	A ou C	M	
B	C ou G ou T	B	
D	A ou G ou T	D	
H	A ou C ou T	H	
V	A ou C ou G	V	



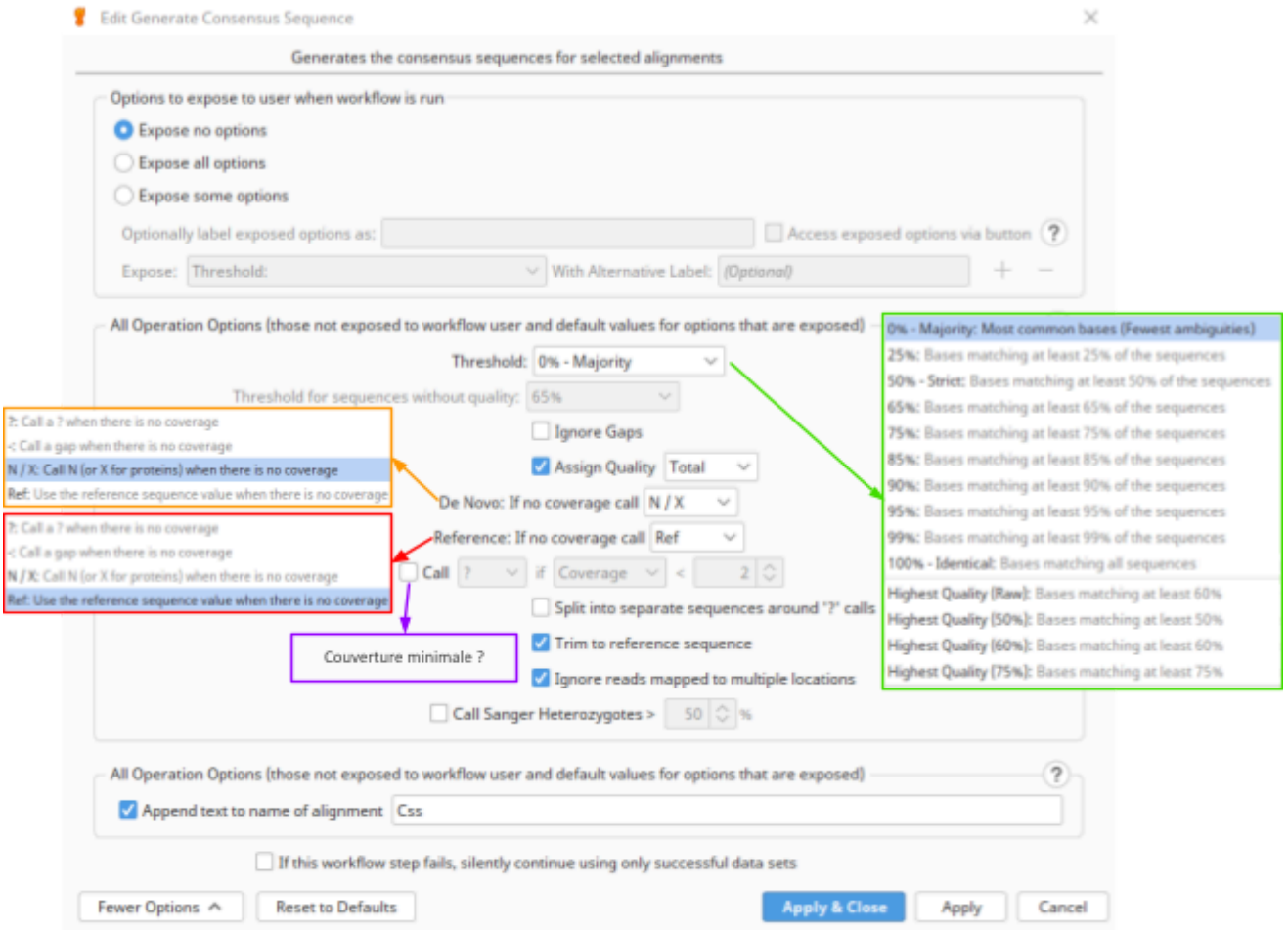
Annexe 8 : Fenêtre de paramètres pour l'alignement sur Geneious Prime



Légende :

NAS = longues lectures NAS
SR = short reads = courtes lectures
Css = consensus
Aligmt = alignement

Annexe 9 : Arbre de décision pour le choix du paramètre en cas d'absence de couverture



Annexe 10 : Fenêtre de paramètres pour la création de consensus sur Geneious Prime

Marie Roynette

Apr 18, 2024, 3:12 AM PDT

Hello Geneious team,

I am using Geneious Prime in the context of my 2 months internship at the INRAE EPGV unit, in France. I solicit your help on a workflow problem.

To explain the situation, I work on the comparison of supposedly similar genes from different plant varieties, starting with long and short sequenced reads of the selected region to get the best consensus possible for each gene of each variety, with the aim of finding polymorphisms between them.

Thus in my workflow, for each variety, I first need to align the long reads on one common reference sequence (the region extracted from the entire reference genome) and generate the consensus. Then, as a polishing for each variety, I need to align its short reads on the consensus generated in the previous step and for each alignment, selecting the consensus corresponding to the right variety as the reference.
For example, by selecting all the files in a folder, I want to align the long reads of variety 1 and 2 on a common reference sequence, and generate consensus 1 and 2. Then, I want the short reads of variety 1 to align on the consensus 1, short reads of variety 2 on the consensus 2 to finally generate a new consensus 1 and a new consensus 2.

However, it seems there is no option available in the workflow panel to select different sequences generated from a previous step as reference sequences for alignments. I search in the model workflows to get inspiration but in vain.

Can you please explain how to do that if Geneious Prime proposes it?

Thank you in advance,

Sincerely

Marie ROYNETTE, INRAE intern

Marie Roynette
Stagiaire BUT
marie.roynette@inrae.fr
Tél. : +33 1 60 87 84 98

Unité EPGV US1279 - Département BAP
Centre Ile-de-France–Versailles-Saclay
Infrastructure de Recherche INRAE GENOMICS
Bat G1 - 2 rue Gaston Crémieux - 91057 Evry Cedex
INRAE EPGV INRAE GENOMICS

Michael Miller (Geneious Prime)

Apr 18, 2024, 5:08 PM PDT

Hi Marie,

Thanks for your email.

This sort of operation isn't preconfigured into Geneious workflows. It might be possible with some customization but it sounds a bit complex for a Geneious workflow. You might consider using Geneious CLI with custom scripts for this.

If you do want to continue designing a workflow, I would suggest starting with the workflow "map reads to reference sequence by name". For this workflow, your sequence list names must either be the same as their reference sequence, or include the reference sequence name as a prefix. This could map all of your long read lists to their reference and create a consensus. Choose to only produce a consensus sequence then you could add a following batch rename step to rename these appropriatly for another iteration of the "map to reference sequence by name". You would need to add a step to pull in the short read lists (selected at the start of the workflow) and then it would map all of these to the consensus sequences from the first interation.

Some other options you might consider:

Spades in Geneious can perform hybrid de novo assemblies when you input long read and short reads together. You could then map these consensus sequences to your references.

Another option would be to map your short reads to your long reads to polish them, then map the polished consensus seqeunces to your reference.

Kind regards,
Michael

Dr. Michael Miller
Support Scientist
help.geneious.com
Geneious

Annexe 11 : Premier échange par mail avec le service support de Geneious Prime

Résumé

Les plantes sont confrontées à diverses maladies et agents pathogènes face auxquels elles ont dû adapter leur réponse immunitaire. Elles ont ainsi développé des résistances spécifiques, dont l'expression est affectée par la variabilité génétique au sein d'une même espèce. La compréhension de ces mécanismes de résistance passe donc par l'étude de leur génome, et notamment des polymorphismes intraspécifiques.
Chez les plantes, les résistances aux maladies et agents pathogènes sont majoritairement contrôlées par des gènes de type NLR, encore mal connus dans leur structure et rôle dans la résistance.
Dans le cadre de sa thèse, Javier BELINCHON MORENO analyse notamment des données de séquençage ciblé longues lectures (technologie ONT NAS) et courtes lectures tout génome de 15 régions génomiques de melon contenant des gènes NLR.

L'étude présentée dans ce rapport complète les travaux de cette thèse en s'intéressant à 6 autres régions génomiques identifiées récemment comme contenant des gènes ou pseudogènes de résistance. La reconstruction automatisée de ces régions d'intérêt avec un workflow Geneious Prime, et l'utilisation d'outils d'annotation tels que BlastX pour localiser des motifs des gènes de résistance, ont permis d'initier l'étude de leurs polymorphismes (SNP et variations de structure). La première analyse par BLASTx de 2 variétés dans la région Chr06:27029991-27032305 a permis de mettre en évidence un motif TIR caractéristique des gènes NLR avec des variations SNP. Dans le cadre de la thèse de Javier BELINCHON MORENO, une annotation expertisée enrichie de l'analyse du polymorphisme de ces régions, permettront d'identifier leurs rôles potentiels dans la résistance aux pathogènes chez le melon.

Abstract

Plants are confronted with various diseases and pathogens to which they have had to adapt their immune response. Thus, they have developed specific resistances. Their expression is affected by genetic variability within the same species. To understand these resistance mechanisms, we need to study their genomes and their intraspecific polymorphisms.
In plants, resistances to diseases and pathogens are mainly controlled by NLR genes, whose structure and role in resistance are still poorly understood.
As part of his thesis, Javier BELINCHON MORENO analyzed targeted long-read (ONT NAS technology) and short-read sequencing data from 15 melon genomic regions containing NLR genes.

The study presented in this report completes the work of this thesis by focusing on 6 other genomic regions recently identified as containing resistance genes or pseudogenes. The automated reconstruction of these regions using a Geneious Prime workflow and the use of annotation tools such as BlastX to locate resistance gene motifs have enabled us to initiate the study of their polymorphisms (SNPs and structural variations). The first BLASTx analysis of 2 varieties in the Chr06:27029991-27032305 region revealed a TIR motif characteristic of NLR genes with SNP variations. As part of Javier BELINCHON MORENO's thesis, expert annotation enriched by polymorphism analysis of these regions will enable the identification of their potential role in pathogen resistance in melons.