Sélection de gènes candidats chez des rosiers des XVIIIe et XIXe siècles et mise en place d'une stratégie de traitement des données permettant l'étude de leur diversité

Nathalie Jacquier

AGROCAMPUS OUEST

☑ **CFR Angers**

☐ **CFR Rennes**

Année universitaire : 2014-2015

Spécialité :

Horticulture

Spécialisation (et option éventuelle) :

Horval

Mémoire de Fin d'Études

☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ d'un autre établissement (étudiant arrivé en M2)

Sélection de gènes candidats chez des rosiers des XVIII^{e} et XIX^{e} siècles et mise en place d'une stratégie de traitement des données permettant l'étude de leur diversité

Par : Nathalie JACQUIER

Soutenu à Angers

Le 07/09/2015

Devant le jury composé de :

Président : Caroline Widehem

Maître de stage : Jérémy Clotault

Enseignant référent : Béatrice Teulat

Autres membres du jury : Charles Eric Durel, examinateur hors formation/juré

Remerciements

Je tiens à remercier Jérémy Clotault pour m'avoir permis de réaliser ce stage au sein de l'équipe, pour son accueil et ses conseils, mais aussi pour sa réactivité lors des imprévus survenus au cours du stage et sa gestion de la communication avec l'EPGV.

Un grand merci à Mathilde avec qui les six mois sont passés très vite et dans la bonne humeur au labo et au bureau.

Merci à Vanessa Soufflet-Freslon sans qui mon CV n'aurait jamais été glissé sur la table de réunion de GDO et qui m'a ouvert les portes de l'IRHS.

Merci à Béatrice Teulat pour avoir accepté d'être ma tutrice durant ses six mois et pour ses conseils quant à la rédaction de ce mémoire.

Je tiens à remercier tout particulièrement l'EPGV pour son accueil et sans qui nos données ne seraient pas ce qu'elles sont !

Je remercie l'ensemble de l'équipe GDO pour son accueil et sa convivialité. Une intention particulière pour Annie avec qui j'ai passé des moments très amusants au labo et Julien, sans qui nos échantillons d'ADN seraient encore incomplets.

Je ne peux bien sûr pas oublier de remercier mes parents pour le temps consacré à la correction de ce mémoire, pour tous les moments partagés entre deux relectures et pour tous ceux déjà devenus mémorables au fil des années !

Je remercie également ma sœur Corinne pour sa persévérance sans faille à tenter de me déconcentrer et pour toutes ses bêtises qui m'ont fait rire.

Un grand merci à Pierre pour la confection de délices blancs pendant la phase de rédaction du mémoire. Cela valait le coup d'attendre trois ans !

Je tiens à remercier tous mes amis proches qui ont le courage de me supporter depuis plus de quatre ans maintenant : Anaëlle avec qui les sorties se comptent par dizaines, Camille pour tous ses gâteaux qui ont ponctué nos soirées, Emelyne pour sa spontanéité et son humour qui ont toujours su animer nos soirées, Justine avec qui les conversations peuvent durer des heures et Dany qui est notre maître d'œuvre et expert informatique officiel. Je remercie aussi les amis que j'ai découverts plus récemment mais qui sont tout aussi importants, en particulier Olivier, Doriane, Benjamin et Yannick. Merci à tous pour votre enthousiasme, votre bonne humeur et votre générosité. Pleins de souvenirs et d'autres à venir !

Enfin, je tiens à faire un clin d'œil à Flippy, Crapolomoche, Catkillerfrog, Little Sugar Plum Fairy, Têtard futé, Poussette et tous les autres qui m'accompagnent depuis des années.

Sommaire

Glossaire

Acide aminé : Unité constitutive des séquences protéiques.

Adaptateur : Petite séquence nucléotidique mise au point de façon à ce qu'elle s'hybride à une autre séquence ou portion de séquence nucléotidique, par complémentarité de séquence.

Allèle : Version d'un gène.

Amorce : Petite séquence nucléotidique de 20 à 30 bases qui est complémentaire d'une portion d'ADN et qui permettra de réaliser une amplification spécifique de l'ADN.

Amplicon : Séquence nucléotidique étant le résultat d'une amplification spécifique par PCR d'une portion d'ADN qui a été bornée par un couple d'amorces.

Blast : Algorithme permettant de détecter des régions similaires entre plusieurs séquences nucléiques (ex : gènes, scaffolds) et/ou séquences protéiques et de les aligner.

Blastn : Blast spécifique lorsque les séquences comparées sont des séquences nucléotidiques.

Cluster (=groupe) : Dans le cas de méthodes de séquençage de 2^{e} génération (Illumina, 454, Ion Torrent, etc.), ensemble de séquences issues de l'amplification d'une séquence initiale unique et qui généreront un read lors de la phase de séquençage. Cette phase de création de clusters permettra de détecter un signal lors du séquençage. Elle est nécessaire pour ces méthodes qui ne sont pas capables de séquencer une molécule unique.

Code cigar: Information issue du mapping indiquant pour un read donné les zones d'insertions, de délétions et correspondance entre la séquence du read et la séquence de référence.

Contigage : Etape qui consiste en l'assemblage de séquences nucléotidiques de petite taille présentant une région de chevauchement, pour aboutir à une plus longue séquence (read contigué). Plus spécifiquement, dans le rapport, on parle de contigage intra-séquence pour l'assemblage des reads forward et reverse d'un même couple pour n'en former qu'un seul. On parle aussi de contigage inter-séquences, pour l'assemblage de contigs issus de couples d'amorces différents d'un même gène.

Dénaturation : Etape durant la PCR permettant aux deux brins d'ADN complémentaires de se séparer à une température élevée.

Dosage d'haplotypes : Comptage du nombre d'haplotypes et de l'occurrence de chacun, afin de connaître le génotype d'un individu.

Epigénétique : Etude des changements héritables (par mitose et/ou méiose) sans changement de la séquence nucléotidique.

Exon : Portion de gène codante qui est conservée après l'épissage/excision et qui sera traduite en séquence protéique.

Extrémité 3' et 5' UTR : Portion de gène en extrémité non codante d'un gène (hors du cadre de lecture). Ces extrémités sont généralement transcrites dans l'ARNm, même si elles peuvent

comporter des introns, mais ne seront pas traduites en séquence protéique. 5’ est en amont du codon initiation et 3’ après le codon stop. Ces régions peuvent servir à la régulation de l’expression des gènes.

Famille multigénétique : Ensemble de gènes, au sein d'un même génome, qui présentent des homologies de séquence et qui sont issus par des phénomènes de duplication anciens d’un gène ancestral commun.

Flag: Information issue du séquençage donnant des informations sur la qualité de l’alignement de chaque read par rapport à une séquence de référence donnée.

Flowcell : Plaque de verre où sont hybridées les séquences nucléotidiques qui seront lues par le séquenceur.

Fluorimétrie : Méthode de dosage utilisant la propriété de fluorescence de certaines molécules comme l’ADN. On mesure la fluorescence qui est proportionnelle à la concentration en ADN.

Gène : Segment d'ADN situé à un endroit bien déterminé d’un chromosome et codant pour une molécule d’ARN fonctionnelle. Il est composé d’introns, d’exons et d’extrémités 3’ et 5’UTR.

Gènes orthologues : Lorsque deux gènes descendent d'un même gène ancestral (on dit qu’ils sont homologues), ils sont orthologues s’il y a eu une évolution séparée des deux gènes après une spéciation, c’est-à-dire après l’apparition d’une nouvelle espèce.

Génome : Ensemble du matériel génétique d'un individu ou d'une espèce.

Génotype : Information portée par le génome d’un organisme qui comprend la composition allélique d’un gène ou de toutes les séquences de cet individu.

Haplotype: Combinaison d’allèles de sites polymorphes présents sur un même chromosome.

Hétérozygotie: Un organisme est hétérozygote pour un gène quand il possède deux (ou plus selon la ploïdie) allèles différents, sinon il est homozygote.

Indels: Mot valise composé d’Insertion et Délétion.

Insert size : Information qui indique le nombre de bases comprises entre les extrémités 5’ de deux reads.

Intron : Portion de gène non codante qui est éliminée après la transcription de l’ADN en ARN lors de l’excision. Elle n’est donc pas traduite en séquence protéique.

Mapping : Alignement optimisé de la séquence d’un read sur une séquence de référence.

Match/mismatch: Match indique qu’une base est identique à celle de la séquence de référence à laquelle elle est comparée. Un mismatch indique une différence entre les bases, c’est-à-dire un SNP. Une portion de séquence nucléotidique matchée indique que cette dernière est totalement alignée sur une séquence de référence. Elle peut contenir des SNP mais pas d’insertions ou de délétions.

Max-poly X : Nombre maximal de succession d'une même base.

MiSeq : Technologie de séquençage mise au point par Illumina®, une entreprise américaine de biotechnologie spécialisée dans le séquençage et le génotypage.

Outgroups : Espèces externes au groupe étudié, permettant de déterminer les allèles ancestraux (communs avec l'outgroup) et les allèles dérivés (différents de l'outgroup) au niveau des sites polymorphes détectés dans le groupe étudié.

PCR : Réaction de Polymérisation en Chaîne. Permet d'obtenir plusieurs copies d'un fragment d'ADN. Dans notre étude, il s'agira d'une **PCR Access Array** mise au point par **Fluidigm®** qui est une compagnie américaine spécialisée dans la conception Integrated fluidic circuit utilisée lors de PCR.

Phénotype : Ensemble des caractères observables d'un individu (ex : couleur des fleurs, parfum, forme des feuilles…).

Phylogénie : Etude des relations de parenté entre individus, populations ou espèces. Elle permet de reconstituer l'évolution des organismes vivants.

Ploïdie : Nombre de lot de chromosomes de base dans le génome. Le rosier a un nombre de base de 7 chromosomes. Un diploïde a donc deux lots de 7 chromosomes, un triploïde en a 3…

Polymorphisme génétique : variations de la séquence nucléotidique d'un gène.

Poolage : Manipulation qui consiste à rassembler différentes données ensemble.

Primer-dimer : Sous-produit de PCR issue de l'hybridation de deux amorces entre-elles.

Quality trimming : Suppression de certaines bases d'un read en fonction de la valeur de qualité de celles-ci.

Read contigués : Dans notre étude, un read contigué sera le résultat du contigage des séquences des reads forward et reverse d'un même couple.

Read : Séquence nucléotidique issue de la lecture des amplicons par un séquenceur.

Remontance : Capacité d'un rosier à fleurir plusieurs fois dans l'année.

Rétrotransposon: Classe d'élément transposable capable de s'insérer à une nouvelle position du génome, par l'intermédiaire d'un ARNm rétrotranscrit.

Run : Utilisation d'un appareil, par exemple de séquençage, et par extension lot de données produites à l'issue de cette utilisation.

Scaffold : Ensemble de contigs orientés et ordonnés issus du séquençage du génome. Dans cette étude, ce sont sur ces scaffolds que se situent les gènes candidats. Un contig est un ensemble de séquences chevauchantes qui peuvent être assemblées en enchaînements plus grands.

Séquençage : Détermination de l'ordre d'enchaînement des nucléotides/résidus constitutifs d'un fragment d'ADN donné.

Séquence de référence : Pour cette étude, séquence nucléotidique d'un gène candidat chez le rosier 'Old Blush' dont le génome est connu.

Séquence nucléique : Succession de nucléotides. Ces séquences sont soit des ADN soit des ARN.

Séquence protéique : Séquence polypeptidique qui constitue les protéines. Elle est le résultat d'une traduction de la séquence nucléique en protéique.

Séquence source : Séquence nucléotidique d'un gène candidat annotée chez une autre espèce que le rosier.

SNP : variation d'une seule paire de bases du génome, entre individus d'une même espèce.

Tm (température de fusion ou demi-dénaturation) : Température pour laquelle 50% de l'ADN est sous forme simple brin et 50% sous la forme double brin.

Transcription : Processus permettant d'obtenir la copie d'une molécule d'ADN en ARN.

Transcriptome : Ensemble des ARN issus de la transcription du génome.

Transcrit : Séquence nucléotidique issue de la transcription de l'ADN.

Workflow: Automatisation d'une succession d'opérations qui permet un traitement de données tout en minimisant l'intervention d'un opérateur.

Zone de chevauchement inter-séquences : Zone commune entre deux reads contigués issus de deux couples d'amorces différents (amorces comprises).

Zone de chevauchement intra-séquence : Zone commune entre le read forward et reverse d'un même couple.

Liste des abréviations

ADN : Acide désoxyribonucléique

EPGV : Etude du Polymorphisme des Génomes Végétaux

GDO : déterminisme Génétique et Diversité des plantes Ornementales

GDR : Genome Database for Rosaceae

Indels : Mot valise Insertion-Délétion

NCBI : National Center for Biotechnology

NGS : Next-Generation Sequencing

PCR : Réaction de Polymérisation en chaîne

SNP : Single Nucleotide Polymorphism

TAIR : The Arabidopsis Information Resource

Liste des figures

Liste des tableaux

Liste des annexes

Introduction

Le rosier est une plante incontournable présente dans la majorité des jardins français mais également visibles sur les terrasses ou les balcons. La rose est en outre la fleur phare de nos bouquets à laquelle s'attache une valeur symbolique, variable selon les coloris des pétales. Elle est sans doute l'une des plantes ornementales la plus cultivée au monde et la plus importante économiquement (Debener et al., 2000). L'an passé dans l'hexagone, elle représentait 2% des achats de végétaux d'extérieur avec près de 7,4 millions de rosiers achetés pour une valeur de 79,7 millions d'euros [1]. Sur le marché de Rungis, plus du tiers des fleurs coupées sont des roses, ce qui reflète sa position de fleur la plus consommée au monde [2].

Son histoire déjà ancienne a laissé de nombreuses traces dans les écrits, notamment à partir du XVIII^{e} siècle où les créations variétales se multiplient. Actuellement, on dénombre plusieurs centaines de nouvelles variétés chaque année [3] dont les innovations de couleur, de forme ou encore de parfum sont primordiales pour qu'elles se démarquent les unes des autres. Les enjeux économiques et son riche passé ont fait du rosier la plante modèle en horticulture ornementale [4].

Les recherches la concernant sont nombreuses et différentes unités françaises mènent des études de thématiques variées. Ainsi, le Laboratoire de Biotechnologies Végétales appliquées aux Plantes Aromatiques et Médicinales de Saint-Etienne étudie les composés volatiles des pétales de rose [5] tandis qu'à Lyon le Laboratoire de Reproduction et Développement des Plantes analyse la morphogénèse florale [6]. A Angers, plusieurs sujets de recherche sont menés au sein de l'Institut de Recherche en Horticulture et Semences (IRHS) où deux équipes travaillent sur le rosier [4]. La première, Arch-E (Biologie Intégrative de l'Interaction Architecture et Environnement), traite des problématiques liées à l'architecture du rosier. Elle vise à améliorer la qualité et la quantité des productions des plantes ornementales en s'intéressant aux facteurs influençant le débourrement des bourgeons ou la mise en place de la ramification. La seconde équipe, GDO (Génétique et Diversité des Plantes Ornementales), au sein de laquelle j'ai travaillé, a différents sujets de recherches. Le premier consiste en l'étude au niveau génétique et moléculaire de caractères d'intérêts qui sont liés à la floraison, l'architecture ou la résistance aux maladies. Le second s'intéresse à l'impact des sélections humaines et naturelles sur la diversité génétique du genre *Rosa* pour mieux comprendre sa structuration et sa gestion. La dernière problématique concerne la mise au point d'une méthodologie pour exploiter les connaissances acquises sur les deux premières thématiques pour la création variétale du rosier et d'autres plantes ornementales.

C'est dans ce cadre d'étude que s'inscrit le projet FloRHiGe [7] qui a pour but d'identifier les facteurs de réussite de l'innovation en horticulture ornementale et plus particulièrement de la sélection du rosier aux XVIII^{e} et XIX^{e} siècles en France, période où la création variétale est foisonnante. Ce projet interdisciplinaire regroupe des généticiens, des botanistes et des bio-informaticiens de l'équipe GDO, ainsi que des historiens du centre de Recherche historique de l'Ouest et du centre François Viète de l'université de Nantes. Des partenariats avec le centre d'Etude du Polymorphisme des Génomes Végétaux d'Evry (EPGV) et quatre roseraies ont aussi été mis en place (roseraie de Nantes, de la Cour de Commer, Loubert et du Val de Marne). Cette mixité permet de répondre à trois grandes problématiques mêlant histoire et génétique qui sont :

1- L'étude de l'évolution des processus d'innovation en lien avec les changements sociétaux et le développement des connaissances.

2- La détermination de l'étendue et de la structure de la diversité génétique des rosiers et l'identification de différents gènes soumis à la sélection et impliqués dans le déterminisme de caractères d'intérêt.
3- La compréhension des facteurs ayant influencé la conservation dans les roseraies et l'impact des méthodes de conservation sur la diversité génétique.

C'est au sein de ce projet FloRHiGe [7] que se déroule la thèse de Mathilde Liorzou qui permettra de répondre à la seconde problématique et en partie à la troisième. L'enjeu sera d'étudier les impacts de la sélection des XVIII^{e} et XIX^{e} siècles et des méthodes de conservation sur la diversité génétique des rosiers. La première partie de sa thèse a été consacrée à l'étude de la diversité du fond génétique des rosiers et de sa structure, ce qui a permis d'identifier 16 groupes génétiques différents. A terme, cette diversité sera comparée à celle de gènes d'intérêt, dits candidats, susceptibles d'avoir été sélectionnés lors des créations variétales des XVIII^{e} et XIX^{e} siècles afin de déterminer une histoire évolutive commune ou parallèle. Cependant, avant toute étude de diversité de ces gènes candidats, plusieurs démarches doivent être entreprises. Ces dernières font l'objet de ce mémoire ainsi que les réflexions menées et les stratégies mise en place pour l'obtention de données analysables.

La première réflexion est portée sur la sélection judicieuse de gènes candidats en raison de leur implication dans l'expression de caractères susceptibles d'avoir été sélectionnés au cours des XVIII^{e} et XIX^{e} siècles. En effet, à cette époque, divers caractères ont été choisis pour l'amélioration des nouvelles variétés de rosiers. Parmi les plus importants, on retrouve la couleur de la fleur, son parfum, le nombre de ses pétales ou encore la remontance, c'est-à-dire la capacité du rosier à refleurir au cours de l'année.

L'analyse de la diversité de gènes candidats s'effectue sur des données de séquençage qui correspondent à la détermination de la succession des bases azotées (A, T, G ou C) présentes au sein d'une séquence d'ADN. Une réflexion en amont du lancement de la production de ces données doit être menée pour déterminer les technologies qui seront employées pour et avant le séquençage. En effet, avant de pouvoir séquencer les gènes candidats, il est nécessaire d'effectuer une amplification de ceux-ci, c'est-à-dire de multiplier le nombre de copies du gène, de façon à atteindre un nombre de copies détectable. Or, les informations désirées sont localisées sur des portions spécifiques du génome, qu'il va falloir identifier puis borner à l'aide de couples d'amorces. La création de ses derniers nécessite une attention particulière.

Enfin, les données obtenues à l'issue du séquençage sont constituées de plusieurs milliers de fichiers qui comportent les séquences des gènes candidats pour un échantillon représentatif de la diversité des rosiers. La quantité d'informations étant très importante, il est nécessaire de mettre en place un traitement bioinformatique des données pour les analyser rapidement et automatiquement.

Nous allons donc présenter la réflexion menée pour le choix de gènes candidats susceptibles d'avoir été sélectionnés chez des rosiers des XVIII^{e} et XIX^{e} siècles et la stratégie de traitement des données mise en place pour permettre l'étude de leur diversité.

Pour répondre à cela, nous fournirons dans un premier temps quelques informations sur les rosiers et sur les gènes candidats, plus particulièrement ceux impliqués dans la couleur des pétales de roses. Nous présenterons ensuite le matériel utilisé pour cette étude et la méthode mise en œuvre pour l'ensemble du projet. Nous expliquerons alors les stratégies mises en place pour la création des amorces et le traitement bioinformatique des données de séquençage. Les résultats obtenus seront alors discutés, critiqués et mis en perspective pour la suite du projet.

Partie 1: Les rosiers

1.1. Quelques informations sur les rosiers

1.1.1. Les rosiers, une longue histoire

Les plus anciennes traces de rosiers, datant de l'Eocène (40 millions d'années), ont été découvertes en Chine dans la région du Liaoning (Guoliang, 2003). Avec l'Europe, ce pays est l'un des berceaux des rosiers (Gudin, 2000). En effet, il accueille une grande source de diversité qui sera exploitée à partir de la fin du XVIII^{e} siècle par l'introduction en Europe des premiers rosiers chinois tels que *R. odorata, R.rugosa* ou *R.wichurana*.

Figure 1 *: Rosa chinensis* 'Old Blush'. *(1998) [8].*

C'est durant ce siècle que la culture des roses va commencer à se multiplier en Europe et ce notamment grâce aux transports maritimes et à l'essor des voyages à visée botanique qui permettront d'importer des rosiers nord-américains ou asiatiques. La popularité de ces plantes pérennes dans les parcs et les jardins de la haute société est alors à son apogée et l'import de diverses espèces permet d'exploiter de nouveaux caractères tels que le parfum ou les couleurs (Meynet, 2001). Apparaît aussi le caractère remontant, c'est-à-dire la capacité du rosier à fleurir plusieurs fois dans l'année, notamment lorsque la rose chinoise *R.chinensis* 'Old Blush' (Fig.1) aurait été introduit en Europe dès 1751, par le botaniste suédois Peter Osbeck (Marriott, 2003). Cependant à cette époque ce sont encore les types européens anciens (Galliques, Damas, Alba, Centifolia, Mousseux) qui sont majoritairement plantés. Cela va changer dès le début du XIX^{e} siècle où le nombre de nouvelles variétés proposées va exploser et se diversifier en combinant ces nouveaux caractères apportés par les rosiers chinois à ceux des rosiers anciens européens plus rustiques (Guoling, 2003). Aujourd'hui, on compte plus de 3000 cultivars disponibles sur le marché (Haudebourg, 1998) et 40 000 variétés recensées [9].

Une évolution des usages a aussi pu être observée, outre sa grande valeur symbolique au fil du temps. En effet, à l'Antiquité le rosier servait essentiellement à des fins médicales (Meynet, 2001). Il fut aussi utilisé pour la fabrication de parfum dont l'odeur est aujourd'hui encore l'une des plus prisée. Actuellement, le rosier est utilisé principalement à des fins esthétiques que ce soit en fleurs coupées ou en plantes d'ornement dans les jardins et les aménagements paysagers. On le cultive aussi en tant que plante en pot et mini-rosiers [1]. D'autres usages se sont développés comme l'usage alimentaire (cynorrhodon, eau de rose) ou pour l'industrie cosmétique (Vukosavljev et al., 2013).

1.1.2 Classification

Les rosiers ont aujourd'hui, suite à de nombreux croisements, une grande diversité phénotypique de par leur port, couleur, forme ou parfum. Leur classification et leurs relations phylogénétiques sont complexes et parfois discutées en raison de difficultés à distinguer les espèces entre elles mais aussi à cause de nombreux hybrides interspécifiques, facilement produits dans la nature. Il est cependant admis que le genre contiendrait au moins 150 espèces.

Parmi les différentes classifications botaniques proposées, l'une des plus usitées est la version de Rehder de 1940, modifiée par Wissemann (Tab.1), (Wissemann, 2003). Celle-ci se compose de 4 sous genres dont *Rosa* qui comprend la majorité des espèces et se divise en 11 sections. De nombreuses classifications horticoles, basées sur l'aspect phénotypique, ont été proposées durant la seconde moitié du XX^{e} siècle dont celle de Jack Hakness en 1989 qui comptait 30 classes différentes (Wisseman, 2003). Ce n'est qu'en 2000 qu'a eu lieu une homogénéisation internationale de la classification du rosier lorsque l'American Rose Society en propose une nouvelle qui prend en compte à la fois les aspects botaniques et les diverses hybridations et sélections (Annexe I) (Cairns 2003). Trois groupes ont ainsi été formés : les espèces botaniques (Tab.1), les roses anciennes (pré-1867, 21 classes) et les rosiers modernes (post-1867, 13 classes). Cette classification est désormais adoptée pour l'enregistrement des nouvelles variétés par la World Federation of Rose Societies (Wisseman, 2003).

Tableau 1 : Classification botanique du rosier de Rheder modifiée par Wissemann. *(Wisseman, 2003).*

Sous genre	Section
Plathyrhodon	
Hesperhodos	
Hulthemia	
Rosa	*Pimpinellifoliae* *Rosa* *Banksianae* *Bracteata* *Caninae* *Carolinae* *Indicae* *Cinnamomeae* *Cassiorhodon* *Synstylae* *Laevigatae*

1.1.3 Des croisements à la base des grands groupes de rosiers

Aux XVIII^{e} et XIX^{e} siècle, le nombre de variétés de rosiers va croître rapidement pour mêler différentes caractéristiques. L'ère des roses modernes va alors débuter à partir de 1867. Cette date « arbitraire » fut choisie car il s'agit de la date d'obtention du 1er hybride de thé 'La France' par le rosiériste Guillot (Fig.2) (Meynet, 2001). On considère qu'entre 10 et 15 espèces du genre auraient été impliquées dans l'obtention des rosiers modernes (Guoliang et al., 2003). Aujourd'hui, la Chine reste une source de diversité puisque près de 82 espèces s'y trouveraient (Joyaux et al., 2001). Les multiples hybridations entre les rosiers asiatiques et européens ont conduits à différents grands groupes horticoles (Annexe II).

Figure 2 : Rosier 'La France'. *(2015) [10].*

Des sélections favorisant la remontance

R. chinensis, R. damascena et R. moschata ont engendré les rosiers thé, appelés ainsi en raison de leur parfum caractéristique. (Testu, 1994). Les croisements entre des rosiers chinois et des rosiers européens (*R. gallica, R. alba, R. damascena*) permettront d'obtenir les « Bourbons », les hybrides de Chine et de Portland. Ces derniers seront à l'origine des « hybrides remontants » tétraploïdes (Meynet, 2011).

Des sélections diversifiant les coloris

A partir du XIX^{e} siècle un élargissement de la palette de couleur va se produire et c'est en 1887 qu'un tournant majeur a lieu lorsque Pernet-Ducher réussit le croisement entre un *Pimpinellifolia* à fleurs jaune très intense et un hybride de Thé. Ce croisement donna le premier tétraploïde jaune *R. foetida* (Fig.3), qui fut présenté en 1897 sous le

Figure 3 : *Rosa foetida* 'Soleil d'or'. *(2015) [11].*

nom de « Soleil d'Or ». Il engendrera la famille des « Pernetiana » et contribuera, par l'apport d'un nouveau pigment (caroténoïde), à la diversification des coloris avec des jaunes francs et de l'orange (Meynet, 2001).

Entre 1922 et 1930 apparut un second pigment coloré orange : la pélargonidine. Celle-ci fut introduite par Kordes dans les hybrides de Thé, notamment avec la variété « Indépendance » dont le succès commercial fut faible mais qui engendra de nombreux descendants (Gallais & Bannerot, 1992).

Des sélections permettant une amélioration de qualités agronomiques

Les rosiers occupent naturellement presque tout l'hémisphère nord. Mais au fil du temps et de l'export des rosiers, les zones de culture ont été élargies, ce qui a amené les hybrideurs à sélectionner des rosiers résistants au froid. Un exemple est l'hybride 'Max Graf' (Fig.4) (*R. wichuraiana* x *R. rugosa)* qui est à la base des rosiers résistants et remontants *R. kordesii* (Meynet J., 2001).

Figure 4 : Rosier 'Max Graf'. *(2004) Rhode Island Rose Society [12]*

1.1.4. Le génome du rosier, son séquençage et ses intérêts à l'avenir…

La première plante ornementale dont le génome est séquencé

Plusieurs dizaines de génomes d'espèces végétales ont déjà été séquencés [13] tels que celui de l'arabette des Dames (*Arabidopsis thaliana*), du pommier (*Malus x domestica)* ou du fraisier (*Fragaria vesca*). Aujourd'hui, c'est au tour du rosier *R.chinensis* 'Old Blush' (*R.chinensis* x *R.odorata* var. *gigantea*) qui devient la première plante ornementale dont le génome est séquencé (550Mb). Ce travail, porté par l'INRA d'Angers et Lyon, est issu d'une collaboration entre des équipes du monde entier [14].

Les annotations structurales et fonctionnelles des gènes seront réalisées et permettront d'obtenir un panorama plus précis de la structure du génome (taille des gènes, familles multigéniques…) [15]. Différents caractères pourront alors être étudiés plus en profondeur bien que les fonctions des gènes devront être vérifiées au fur et à mesure des recherches d'équipes (régulation de l'expression des gènes, épigénétique…) [16].

Les diverses ploïdies des rosiers

Chez les rosiers, les niveaux de ploïdie diffèrent entre, et parfois au sein, des différentes espèces. Le nombre de base de chromosomes est x=7. Les individus peuvent être de diploïde à décaploïde (*R. praelucens* Byhouwer, (Jian et al., 2010)). Plus la ploïdie augmente, plus le nombre d'espèces concernées diminue. Cette ploïdie augmente selon un gradient Sud-Nord avec la rudesse du climat. (Krussmann, 1981). Beaucoup de triploïdes sont apparus après les croisements entre les rosiers européens (4x=28) et les chinois (2x=14). La plupart des rosiers modernes sont triploïdes ou tétraploïdes, les thés chinois et Noisette sont souvent diploïdes tandis que les Gallica, Centifolia et Damascena sont tétraploïdes (Crane & Byrne, 2003). Durant notre étude, il sera important de connaître la ploïdie des rosiers étudiés, plus particulièrement pour la réalisation du dosage d'haplotypes dont nous reparlerons ultérieurement (partie 3.4). Cela permettra notamment de déterminer le nombre maximal d'allèles possible pour chacun des individus étudiés.

1.2. Les gènes candidats sélectionnés

Pour l'étude de la diversité, dix gènes candidats ont été retenus pour leur intérêt dans la sélection des rosiers. Ce choix a été réalisé soit à partir d'un travail bibliographique pour les gènes candidats de la couleur, soit à l'aide du partage d'informations entre des équipes de recherche.

Dans le cas des gènes impliqués dans la couleur, nous allons voir dans un premier temps quelle est l'origine de la couleur et quels pigments en sont responsables (les caroténoïdes ne seront pas inclus). Nous expliquerons alors comment ces derniers sont formés et quels sont les trois gènes clefs qui ont été retenus pour l'étude. Des informations bibliographiques permettront d'apporter des informations sur chacun d'eux. Enfin, nous présenterons les autres gènes candidats retenus.

1.2.1. L'origine de la couleur des roses et les facteurs l'influençant

La coloration des fleurs est due à trois grandes familles de molécules : les flavonoïdes, les caroténoïdes et les chlorophylles. Chez les rosiers, ce sont principalement les anthocyanes (pigment flavonoïques) et les caroténoïdes qui sont responsables de la couleur des fleurs. Cependant, les chlorophylles sont responsables des pétales verts ou blancs et vert chez de rares espèces (ex : *Rosa* 'Super Green', 'Amandine', 'Wimbledon' [17]), mais ils ne feront pas l'objet de notre étude.

Chez le rosier, la couleur dépend, en plus de facteurs génétiques, de plusieurs paramètres physiologiques ou environnementaux (Schulz, 2003):

- la concentration en différentes anthocyanes : en général et selon la composition en pigments au sein du pétale, plus elle est élevée, plus la couleur est sombre.
- la forme des cellules : les cellules coniques augmentent la proportion de lumière incidente ce qui accroît l'absorption par les pigments et conduit à des pétales plus sombres et à une saturation de la couleur. L'effet inverse s'observe pour des cellules plates (qui sont favorisées à températures élevées) (Zhao & Tao, 2015).
- la présence de copigments : incolores, ils s'associent aux pigments et permettent de modifier leurs propriétés d'absorption, d'intensifier et fixer les couleurs. (Fukuchi-Mizutani et al., 2011). Les pétales de roses ont la particularité de ne contenir que des flavonols, qui s'avèrent être des copigments moins efficaces que les flavones qui ne peuvent être produits par les rosiers (Katsumoto et al., 2007).
- le pH vacuolaire : celui des cellules épidermiques des pétales de rose est faible mais variable selon les espèces et cultivars (entre 3,69 et 5,78) (Katsumoto et al., 2007).
- la température : élevée, elle favorise les pétales aux couleurs claires car certains gènes de la voie de biosynthèse des pigments (*DFR* ou *F3'H*) sont sous-exprimés. Il y a donc moins de pigments stockés (Fig.5). A l'inverse, les températures faibles favorisent des couleurs plus sombres. (Zhao & Tao, 2015)
- L'eau : un stress hydrique conduit à une accumulation des anthocyanes mais s'il se prolonge on peut observer l'effet inverse (Zhao & Tao, 2015).
- La lumière : intensité, qualité, photopériode modifient l'intensité des couleurs.

1.2.2. La voie de biosynthèse des anthocyanes

La voie de biosynthèse des anthocyanes permet d'obtenir des pigments colorés, les anthocyanes (ou anthocyanines) mais elle est aussi à la base de la synthèse de copigments, les flavonols. Certains gènes de cette voie de biosynthèse sont encore mal connus. En effet, ils ont parfois été récemment découverts, comme le gène *RhGT1* qui est spécifique aux rosiers, (Ogata et al., 2005) ou ils font partie de familles multigéniques, comme la CHS, dont tous les gènes ne

sont pas encore identifiés (Forkmann, 2003). Les gènes, notamment ceux concernant les dernières étapes de la biosynthèse, sont contrôlés de façon spécifique et ce contrôle serait différent selon les cultivars (Fukuchi-Mizutani et al.,2011). Ils peuvent être activés ou non selon les stades de développement de la plante ou les facteurs environnementaux. (Jay et al., 2003). De plus, les transcrits peuvent être différents entre les feuilles et les pétales (Forkmann, 2003).

Les rosiers contiennent 2 anthocyanes principales qui sont stockées dans la vacuole des cellules de l'épiderme supérieur du pétale : la cyanidine et la pélargonidine. Chez les cultivars plus sombres, elles se trouvent aussi parfois dans le parenchyme palissadique et l'épiderme inférieur (Jay et al., 2003). Il faut noter que les cyanidines et ses dérivés sont les pigments majoritaire au sein des pétales des rosiers (de l'ordre de 80%) hormis chez les Banksia, les Caninae, les Pimpinellifolia et les Hybrides perpétuels (Fukuchi-Mizutani et al., 2011).

La présence de péonidine, un dérivé de la cyanidine, est détectée chez certains rosiers comme *R.rugosa* (De Vries et al., 1974). En revanche, la delphinidine n'est pas présente à l'état naturel. En effet, elle fut introduite par transgénèse par l'ajout du gène *F3'5'H* afin obtenir des roses bleues (Katsumoto et al., 2007). De même, certaines molécules comme les chalcones, qui expliquent la couleur jaune chez certaines espèces telles que le pétunia, ne peuvent pas être accumulées (De Vries et al., 1974) chez le rosier. Le schéma de biosynthèse des anthocyanes est décrit ci-dessous (Fig.5) (Forkmann, 2003) (Ogata, 2005).

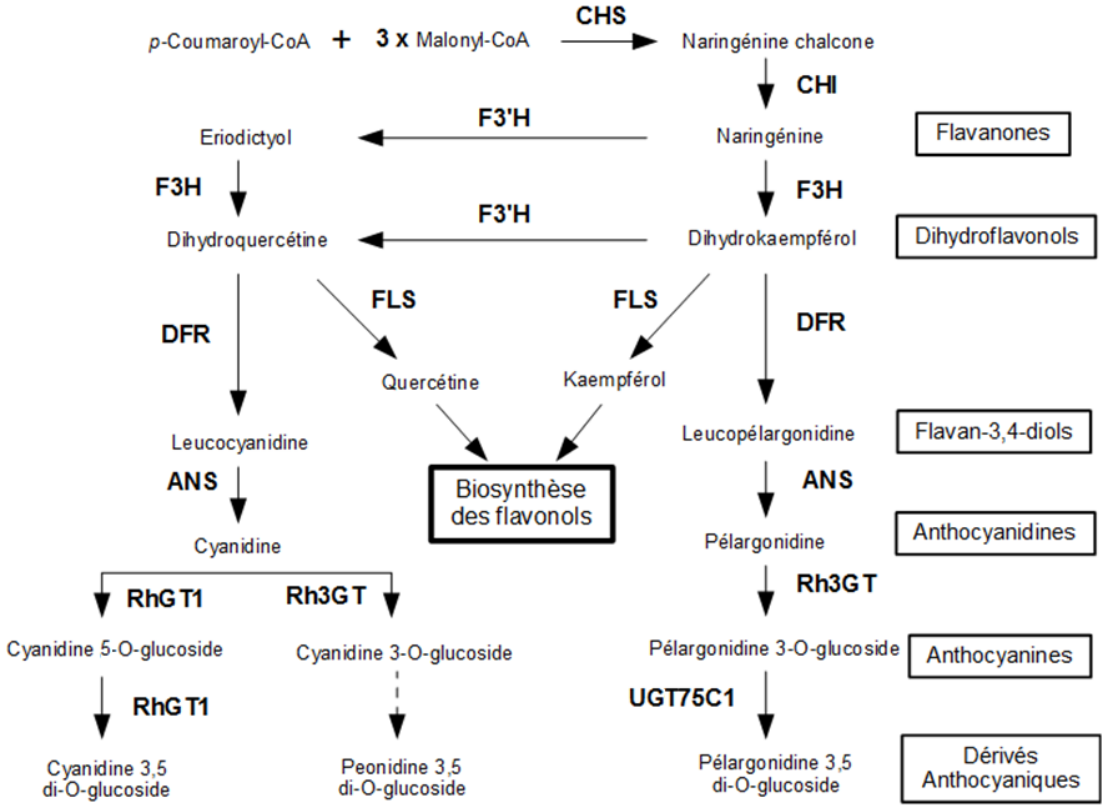


Figure 5 : Voie de biosynthèse des anthocyanes et de ses dérivés chez le rosier. *Les enzymes impliquées sont :* CHS=*Chalcone synthase,* CHI=*Chalcone isomérase,* F3H=*Flavanone synthase,* F3'H=*Flavonoïdes 3' hydroxylase,* FLS=*Flavonol synthase,* DFR=*Dihydroflavonol 4-reductase,* ANS=*Anthocyanidine synthase,* RhGT1=*Anthocyanidin 5,3-glucosyltransférase,* Rh3GT=*Glucosyl transferase,* UGT75C1=*Anthocyanidin 3, 5-glucosyltransferase (Forkmann,2003 et Ogata, 2005).*

Une composition en pigment variable selon la couleur des pétales

Chez les rosiers, les nombreux croisements interspécifiques ont donné lieu à une grande variation de coloris des fleurs. Ainsi, les roses anciennes européennes (galliques, Alba, Centifolia, Damas, Mousseux) ont des coloris allant du blanc au rouge en passant par le rose et

occasionnellement du jaune pâle. (Schulz, 2003) La quercétine et le kampférol sont présents dans l'ensemble des pétales y compris dans les blancs où ils n'interagissent avec aucun autre pigment [18]. La pélargonidine, qui n'est jamais présente sans cyanidine, est dominante dans les pétales orange mais pas dans ceux avec un peu de jaune, rose ou rouge [18].

Les pigments présents selon les coloris sont :

- **Rouge et rose :** Ces coloris sont liés à la présence de cyanidine et de ses dérivés qui peuvent être couplés ou non à la pélargonidine dans le cas des rouges vifs. On trouve aussi dans ce dernier cas des caroténoïdes ((De Vries, 1973). Les rosiers rouges et sombres n'ont pas beaucoup de flavonols et ont un pH faible. (Katsumoto et al., 2007).
- **Blanc et jaune :** Les variétés blanches ou jaunes ne contiennent pas d'anthocyanidines, qu'elles ont l'incapacité de stocker dans leurs pétales (Katsumoto et al., 2007). De plus, les purs blancs ne contiendraient que des flavonols (De Vries, 1973**)** et seraient dus à un blocage en début de la voie de biosynthèse (Forkmann, 2003). On distingue deux types de jaunes : Le jaune pâle, présent avant l'introduction des variétés chinoises et le jaune franc, plus sombre, introduit en 1920 par *R. foetida* 'Soleil d'or' duquel découlent de nombreuses variétés. Ce dernier est dû à la présence des caroténoïdes (Meynet, 2001).
- **Violet :** Il est dû aux cyanines et parfois de son dérivé, la péonine (Jay et al., 2003).
- **Orange :** Les couleurs oranges tout comme certains roses ou rouges brillants viennent de dérivés de la pélargonidine et parfois des caroténoïdes [18].

1.2.3. Les trois gènes candidats retenus

DFR ou Dihydroflavonols Reductase:

Le gène *DFR* a un rôle clef en début de voie de biosynthèse des anthocyanes. En effet, il code pour une enzyme du même nom qui a la particularité de reconnaître deux substrats : la dihydroquercetine (DHQ) qui est un précurseur de la cyanidine, et le dihydrokaempférol (DHK), qui est un précurseur de la pélargonidine. Les rosiers composés de pélargonidine ont une DFR qui peut utiliser à la fois le DHK et la DHQ tandis que ceux avec uniquement des dérivés de cyanidine ont une DFR dont le substrat spécifique est le DHQ. Il a ainsi été établi l'existence de deux enzymes différentes **(Schulz, 2003)**.

De plus, il a été mis en évidence grâce à une étude sur le pétunia que la spécificité de la DFR est due à la modification d'un, ou plusieurs, acide(s) aminé(s) dans une zone clef de la séquence du gène. Cette zone clef de 26 acides aminés déterminerait la capacité à réduire le DHK qui peut varier entre différentes espèces mais aussi entre cultivars de la même espèce et donc modifier la couleur des fleurs **(Johnson et al., 2001).** La DFR du pétunia n'est pas spécifique du DHK contrairement à celle du rosier. (Fig.6).

Figure 6: Comparaison de séquences protéiques de DFR des différentes espèces. *De haut en bas, alignement des séquences protéiques de l'enzyme DFR du Gerbera, Rosier, muflier, oeillet et maïs acceptant la DHK comme substrat et du pétunia qui ne l'accepte pas. En gras les acides aminés communs, les * indiquent un changement par rapport à la séquence protéique du pétunia (Johnson et al., 2001).*

1.2. Les gènes candidats sélectionnés

La mutation de la valine en leucine V133L, de l'asparagine en acide aspartique N134D et de l'acide glutamique en glutamine E145Q auraient un rôle clef dans la zone de spécificité du substrat que nous avons pu localiser au sein de la séquence protéique (Fig.7) et nucléique (Fig.8) du rosier.

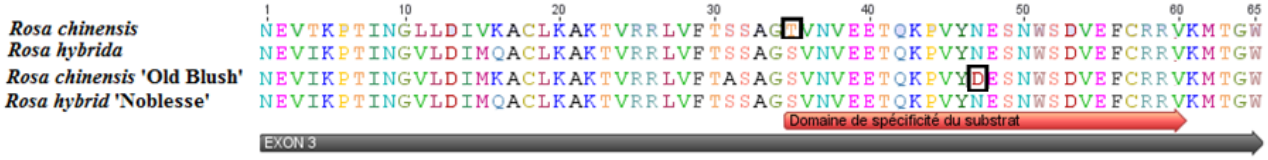


Figure 7 : Comparaison de 4 séquences protéiques de DFR de 4 cultivars de rosiers. *Elles sont issues de la traduction de l'exon 3 du gène DFR et 2 mutations sont observables dans les carrés noirs. La flèche rouge indique la zone de spécificité du substrat identifiée dans l'étude de Johnson et al., 2001. (2015) N. Jacquier*

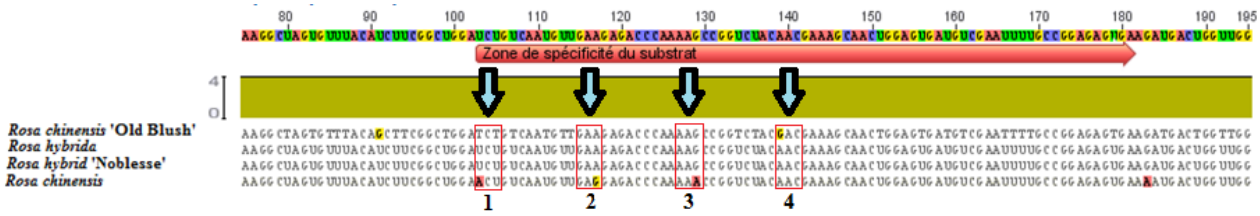


Figure 8 : Comparaison d'une partie de la séquence nucléotidique de l'exon 3 de *DFR* de 4 cultivars de rosiers. *Cette portion de séquence correspond à la zone de spécificité du substrat identifiée précédemment. 4 SNP sont identifiables et les triplets de nucléotides (codons) sont indiqués par des flèches et des encadrés rouges. (2015) N. Jacquier*

Grâce à la comparaison entre les mutations au sein des séquences protéiques (Fig.7) et les SNP présents au sein des séquences nucléiques (Fig.8), il a été possible de déterminer quels changements de base peuvent entraîner une modification d'acide aminé. Ainsi, en raison de la redondance du code génétique (différents triplets de nucléotides, appelés codons, code un même acide aminé), certaines mutations n'auront pas d'impact sur la traduction des séquences protéiques en nucléiques. Dans le cas de du gène *DFR*, les 2ème et 3ème codons (2 et 3 sur Fig.8) codent réciproquement pour l'acide glutamique et la lysine, la présence du SNP n'ayant pas d'impact du fait de la redondance. En revanche, le SNP au sein du 1er codon entraîne un changement d'acide aminé puisque la sérine codée par le triplet TCT devient ACT pour *Rosa chinensis* et l'acide aminé codé est donc la thréonine. Il en va de même pour le 4ème codon où le SNP change le triplet AAC codant pour l'asparagine en GAC codant pour l'acide aspartique. Il sera intéressant de mener une analyse spécifique de cette zone de la DFR à l'issue du traitement des données de diversité.

De plus, il a été observé que le pic d'expression de la DFR correspond au pic d'accumulation des anthocyanines. Ce taux diminue quand les pétales sont complètement ouverts (Suzuki et al., 2014). Le taux de DFR dans les pétales est régulé de manière transcriptionnelle au cours du développement et ce en parallèle de la production d'anthocyanines. Il est faible dans les premiers stades du développement floral puis augmente lorsque les pétales sont colorés et que les sépales commencent à s'ouvrir (Tanaka et al., 1995).

Ce résultat ainsi que l'étude de la zone de spécificité du substrat de la DFR ont fait que ce gène a été retenu car chez le rosier il pourrait expliquer différentes nuances et couleurs du pétale en influençant sur la présence ou la quantité de cyanidine ou pélargonidine.

1.2. Les gènes candidats sélectionnés

F3'H ou Flavonoïdes 3' hydroxylase :

Le gène *F3'H* code pour la F3'H qui est un précurseur de la voie de biosynthèse des anthocyanes. Cette enzyme permettrait également d'obtenir de l'eriodictyol à partir de la naringénine chez certains rosiers (Schulz, 2003). Une mutation spontanée au sein de la séquence du gène serait apparue tardivement (De Vries et al., 1974) et aurait entraîné la synthèse d'une enzyme moins efficace permettant ainsi la synthèse de pélargonidine.

De plus, au sein des rosiers, les différents ratios Kaempférol/Quercétine et Pélargonine/Cyanidine seraient probablement dus à une activité plus ou moins importante de la F3'H dont l'extrémité C-terminale du gène serait très importante pour la réaction enzymatique (Tanaka & Brugliera, 2006).

FLS ou Flavonol synthase :

La FLS est une enzyme qui permet d'obtenir deux flavonols à la base de la voie de biosynthèse des co-pigments : le kaempférol et de la quercétine. La FLS des roses, surtout exprimée aux premiers stades du développement floral, peut utiliser plusieurs dihydroflavonols qui sont : le dihydrokaempférol et la dihydroquercétine (Suzuki et al., 2014). Ce gène a été retenu en raison de sa position clef qui pourrait influencer la présence de certains co-pigments ou une variation de leur ratio. De plus, il est possible que des différences au sein de la séquence nucléique du gène dans une zone permettant la fixation du substrat modifient la composition en co-pigments des pétales et influent donc sur leurs couleurs ou leurs nuances.

1.2.4. Les autres gènes candidats retenus

Les autres gènes candidats retenus pour l'étude concernent différents caractères d'intérêt:

- L'architecture avec l'étude du gène *BRC1* qui est impliqué dans le développement des bourgeons par arrêt de leur croissance (Aguilar-Martínez et al., 2007).

- La floraison avec la sélection de 3 gènes candidats : *KSN*, pour son rôle dans la remontée de floraison (Itawa et al., 2012), *FT* qui modifie la date de floraison (Kobayashi et al., 1999) et *AGAMOUS* qui détermine le nombre de pétales (Foucher et al., 2008).

- Le parfum (Roccia, 2013) avec des gènes impliqués dans la synthèse de molécules odorantes qui sont *NUDX* pour la synthèse des monoterpènes (Magnard et al., 2015) et *PAAS* pour celle du phényléthanol (Machenaud, 2010).

- La couleur due à la présence de caroténoïdes et pour laquelle *ZDS* est un des gènes clefs de la voie de biosynthèse de ces pigments orange (Zhu et al., 2010).

Partie 2 : Matériels et Méthodes

Le schéma ci-dessous (Fig.9) permet de replacer les différentes étapes réalisées au cours de l'étude et de montrer celles qui ont été externalisées au centre d'Etude du Polymorphisme des Génomes Végétaux d'Evry (EPGV).

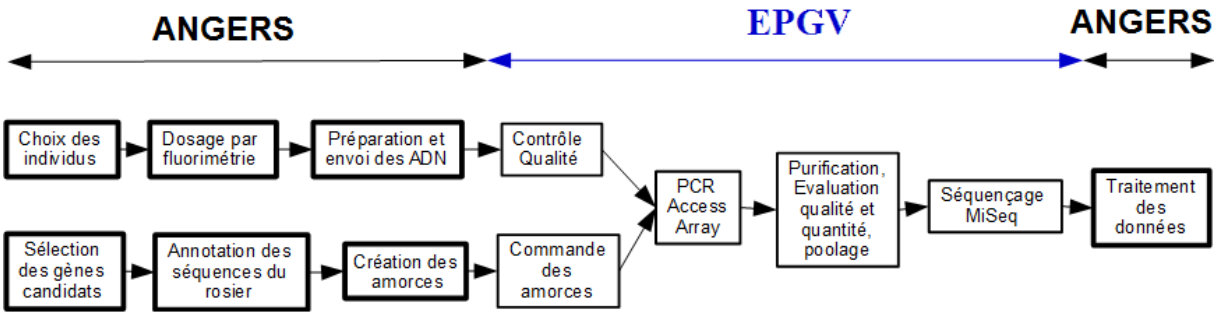


Figure 9 : Schéma bilan des différentes étapes réalisées durant le stage *(2015). N. Jacquier*

Les étapes concernant la création des amorces et le traitement des données seront développés dans la partie résultat (partie 3).

2.1. Matériels

2.1.1. Matériel biologique

La sélection de 8 « individus test »

Les « individus test » ont été utilisés pour vérifier d'une part le bon fonctionnement de la PCR, des amorces et des machines, d'autre part pour déterminer la concentration minimale en ADN nécessaire pour une bonne amplification.

Pour ce faire, les 8 échantillons envoyés avaient (Tab.2) :

- Un volume de 119 µL
- Une concentration supérieure à 15 ng/µL, mesurée par absorbance avec le Nanodrop ND 1000®.
- Différents niveaux de ploïdie pour déterminer si les concentrations devaient être adaptées en fonction de cela.

Les « individus test » ont été utilisés à différentes concentrations : 2, 5 et 10 ng/µL.

Tableau 2 : Récapitulatif des échantillons des 8 individus tests

Accession	Date d'extraction	Concentration en ng/µL (Dosage Nanodrop® à Angers)	Ploïdie	Utilisation à la concentration (Dosage au Qubit® par l'EPGV)		
				2 ng/µL	5 ng/µL	10 ng/µL
directeur_alphand	10/01/2014	15,30	4x	X	X	
ferdinand_chaffolte	10/01/2014	127,40	2x		X	
old_blush	Non connue	423,90	2x		X	X
elisabeth_d_angleterre	10/01/2014	24,60	5x		X	X
rosa_brunonii	14/01/2014	82,70	2x		X	X
marjolin	15/01/2014	110,70	4x		X	X
rosa_xanthina	14/01/2014	144,50	2x		X	X
ville_de_toulouse	15/01/2014	43,90	4x		X	X

Les 48 couples d'amorces utilisés dans ce premier « run test » concernent ceux mis au point pour amplifier les gènes : *DFR* (6 couples), *F3'H* (5 couples), *FLS* (13 couples), *ZDS* (13 couples) et *AGAMOUS* (8 couples sur 14). La *CHS* (3 couples) fut retirée de l'étude par la suite

car ce gène appartient à une famille multigénique mal connue. Pour les run suivants, les couples utilisés sont ceux permettant d'amplifier : *DFR* (6 couples), *F3'H* (5 couples), *NUDX* (13 couples), *ZDS* (13 couples), *KSN* (4 couples), *FT* (5 couples), *PAAS* (3 couples) et *AGAMOUS* (14 couples).

Le choix de la population de rosiers

Le choix des 384 rosiers de l'ensemble de l'étude a été réalisé par les généticiens, les botanistes et les historiens et ce grâce à plusieurs critères.

- Les rosiers ayant été phénotypés ont été choisis de façon préférentielle (308 individus). Cela permettra ainsi de réaliser une comparaison phénotypes/génotypes par génétique d'association.
- Les 16 groupes génétiques déterminés lors de la première partie de la thèse de Mathilde Liorzou ont été représentés façon équilibrée. On évite ainsi le risque sur-représentation d'un groupe par rapport à un autre.
- Les dates d'obtention des variétés ont été représentées de manière équilibrée et proportionnelle par rapport à l'ensemble des rosiers disponibles.
- Suite à la lecture d'un corpus d'archives et de documents du XIXe siècle, analysé par les historiens, les rosiers remarquables et/ou ayant beaucoup de descendants ont été sélectionnés pour leur intérêt historique et leur importance dans l'obtention des nouvelles variétés.
- Enfin, pour des raisons pratiques et un gain de temps, les rosiers pour lesquels de l'ADN ou des feuilles lyophilisées étaient disponibles dans nos stocks ont été privilégiés. De même, l'accès à du matériel frais devait être facilement réalisé en cas de ré-échantillonnage.

Pour l'étude de la diversité, 5 répétitions de Old Blush seront réalisées et 3 outgroups (espèces externes au groupe étudié) ont aussi été retenus (*Sanguisorba angustifolia*, *Geum urbanum* et *Potentilla reptens*). Ces espèces ont été choisies en raison de leur proximité avec le genre *Rosa*. Elles permettront de déterminer quel est l'allèle ancestral, qui est partagé par l'ensemble des outgroups et quel est l'allèle dérivé, qui est différent des outgroups et est apparu depuis la divergence avec le genre *Rosa*.

Les conditions d'envoi des 384 échantillons

L'ADN des échantillons sélectionnés pour l'étude ont été envoyés à la plateforme d'Evry en respectant certaines conditions :

- Un volume entre 35 et 50 µL, le volume minimal étant exigé par Fluidigm® pour l'amplification pour faciliter le pipetage par robot.
- Une concentration minimale de 10 ng/µL déterminée à partir de nos échantillons tests et mesurée par fluorimétrie au Hoescht (Annexe III). Ce seuil est la limite permettant l'obtention de suffisamment de copies d'une portion d'ADN, appelées amplicons. Ces derniers seront utilisés pour le séquençage et l'obtention de reads analysables, c'est-à-dire les séquences issues du séquenceur.

2.1.2. Logiciels

Les logiciels utilisés durant notre étude sont :

- Geneious Pro 5.3.6© pour toutes les phases allant de la sélection à la création des amorces (Kearse et al., 2012), [19].
- L'instance Galaxy ABiMS de la station biologique de Roscoff pour la partie de traitement bioinformatique allant de la conversion des données brutes issues du

séquencage MiSeq au mapping (Blankenberg et al., 2010), (Giardine et al., 2005), [20].

- Microsoft® office Excel® 2007 pour la partie de contigage, filtrage des erreurs de séquençage et dosage des haplotypes.

2.1.3. Bases de données utilisées pour l'obtention des « séquences sources » des gènes candidats

Les « séquences sources » et les « séquences de référence »

Avant la création des amorces, il est nécessaire d'annoter les gènes candidats chez le rosier de référence *Rosa chinensis* 'Old Blush'. Ceci consiste en l'identification sur la séquence d'ADN (Fig. 10) des introns (portions de gènes non codantes), des exons (portions de gènes codantes et traduites en séquence protéique qui constituent les protéines) et des extrémités 5' et 3'UTR (souvent utiles pour la régulation de l'expression des gènes).

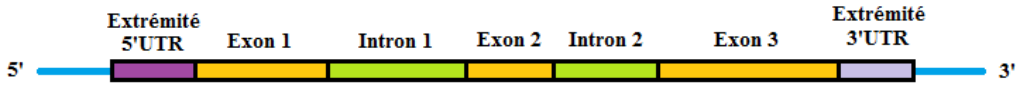


Figure 10 : Structure d'un gène *(2015) N. Jacquier*

Pour effectuer cela, il faut au préalable sélectionner une séquence connue et déjà annotée du gène candidat chez une espèce proche. Cette séquence sera alors comparée à la séquence du rosier et servira de base de travail pour l'annotation. Nous appellerons par la suite cette séquence « séquence source » pour ne pas confondre avec la séquence qui sera annotée chez le rosier et qui sera notre « séquence de référence ».

Les bases de données utilisées pour l'obtention des « séquences sources » annotées

Pour certains de nos gènes candidats, des membres de l'équipe GDO ou d'autres équipes collaboratrices avaient déjà annoté différents gènes chez le rosier et nous ont communiqué les séquences de référence. Pour les autres gènes, il a fallu rechercher puis récupérer une/des séquence(s) sources déjà annoté(es) chez le rosier ou bien chez des espèces voisines les plus proches possible de celui-ci. Ainsi, le fraisier a été choisi préférentiellement en raison de sa proximité phylogénétique avec le rosier, suivi par le pommier et *Arabidopsis* pour lesquels les trois génomes sont disponibles. Pour ce travail, trois grandes bases de données sont disponibles pour obtenir des séquences sources (Annexe IV) :

- **GenBank®** : Ensemble de données comprenant la banque d'ADN du Japon (DDBJ= DNA DataBank of Japan), le laboratoire Européen de biologie moléculaire (EMBL=European Molecular Biology Laboratory) et la base GenBank du NCBI (National Center for Biotechnology Information) [21].
- **TAIR** (The Arabidopsis Information Resource) : Base génétique et moléculaire d'*Arabidopsis* [22].
- **GDR** (Genome Database for Rosaceae) : Base de données des Rosacées [23].

2.2. Méthodes

2.2.1. Méthode de sélection des séquences sources

La sélection de la séquence de base de travail et l'obtention des scaffolds

De façon générale, il était intéressant de sélectionner des séquences ayant été travaillées par une équipe de recherche plutôt qu'issues d'une annotation automatique comme c'est le cas du génome du Fraisier (*Fragaria vesca*). En effet, le gène a ainsi déjà été étudié et annoté avec soin et de façon plus précise par l'équipe qui a souvent obtenu des transcrits facilitant cette

annotation. De plus, les résultats obtenus ont généralement conduit à une publication à l'aide de laquelle il est possible d'obtenir des informations complémentaires.

De plus, lorsque plusieurs séquences sources étaient disponibles, la plus proche du rosier a été retenue. Celle-ci se définit comme telle lorsque :

- Elle 'blaste' avec le génome du rosier, c'est-à-dire que la séquence s'aligne sur une portion d'un ou plusieurs scaffolds de 'Old Blush'. Ces portions correspondent au gène candidat qui est un orthologue de la séquence source.
- L'ensemble des exons de la séquence source est aligné le long du scaffold.

Grâce à cet alignement, on peut alors par comparaison de séquences repérer les positions des introns, des exons et des extrémités 5' et 3'UTR.

Méthode de confirmation des annotations

Cette étape ne doit pas être négligée, car cela permet de confirmer la distinction entre les exons et les introns, ce qui est essentiel pour le design des amorces et l'optimisation du choix de celles-ci, comme nous le verrons par la suite (partie 3.1).

Pour confirmer notre annotation, plusieurs informations ont été exploitées :

- Des données bibliographiques sur le rosier ont parfois permis d'obtenir des informations sur certains gènes candidats telles que le nombre et la longueur des exons ou la longueur du gène.

- L'obtention de la séquence chez le rosier de certains transcrits des gènes candidats, c'est-à-dire le produit issu de la transcription de la séquence d'ADN en ARN. Il faut préciser que les introns ne sont plus présents au sein de la séquence des transcrits et donc l'alignement avec les scaffolds est discontinu. Ces séquences sont disponibles dans les bases de données du GDR ou du transcriptome du rosier Toulouse [24]. De plus, lorsque l'on a trouvé plusieurs séquences sources, elles ont pu être utilisées pour confirmer l'annotation des scaffolds.

2.2.2. Manipulations de laboratoire réalisées à Angers

Extraction de l'ADN et dosage par fluorimétrie

Nous avons extrait les ADN de rosier à l'aide du QIAGEN® DNeasy® 96 Plant Kit à partir de feuilles de rosier lyophilisées ou fraîches. Seules les feuilles de 'Old Blush' étaient congelées à -20°C. (protocole en annexe V). Le dosage des ADN a été réalisé par fluorimétrie au Hoescht en (annexe III). L'appareil de mesure est le TECAN Infinite ® 200.

2.2.3. Externalisation de l'amplification et du séquençage.

La PCR Access Array IFC de Fluidigm® : l'obtention d'amplicons

La PCR, ou Réaction de Polymérisation en Chaîne, permet d'obtenir pour un couple d'amorces donné, plusieurs copies de la séquence d'ADN ainsi bornée, appelés amplicons. Dans notre étude, l'amplification a été réalisée par PCR Access Array de Fluidigm® à 4 amorces dans une puce microfluidique, la plaque 48.48 Access Array IFC (Integrated Fluidic Circuit) [25]. Cette technologie permet de réaliser 2304 réactions (48 individus*48 couples d'amorces) rapidement et dans un volume de réaction très réduit (35 nL). De plus, cette technologie permet de réduire le nombre et le temps des manipulations et a l'avantage de s'adapter à différents appareils de séquençage du type Illumina, IonTorrent ou encore Roche 454 [25].

2.2. Méthodes

Cette amplification se divise en 3 grandes étapes : Pré-PCR, PCR et post-PCR (Fig.11).

Figure 11 :Trois appareils différents utilisés pour la pré-PCR, la PCR, et la post-PCR *[25]*

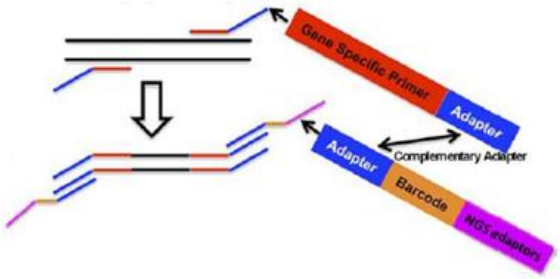


Figure 12 : Les 4 amorces utilisées durant le PCR Access Array *(image fournie par l'EPGV)*

Quatre amorces différentes (Fig.12) sont utilisées pour permettre d'amplifier des zones spécifiques de l'ADN et d'identifier de façon unique chacune des réactions :

- Les amorces spécifiques (gene specific primer), que nous avons mises au point et qui sont uniques pour un couple d'amorces donné.
- Les adaptateurs universels, qui permettent d'ajouter les codes-barres et les adaptateurs NGS.
- Les codes-barres, qui permettent d'identifier chaque individu.
- Les adaptateurs NGS, qui sont spécifiques selon la méthode de séquençage utilisée ultérieurement, dans notre cas MiSeq.

Le schéma général de la PCR est le suivant (Fig.13), le protocole complet est disponible sur le site internet de Fuidigm® [26].

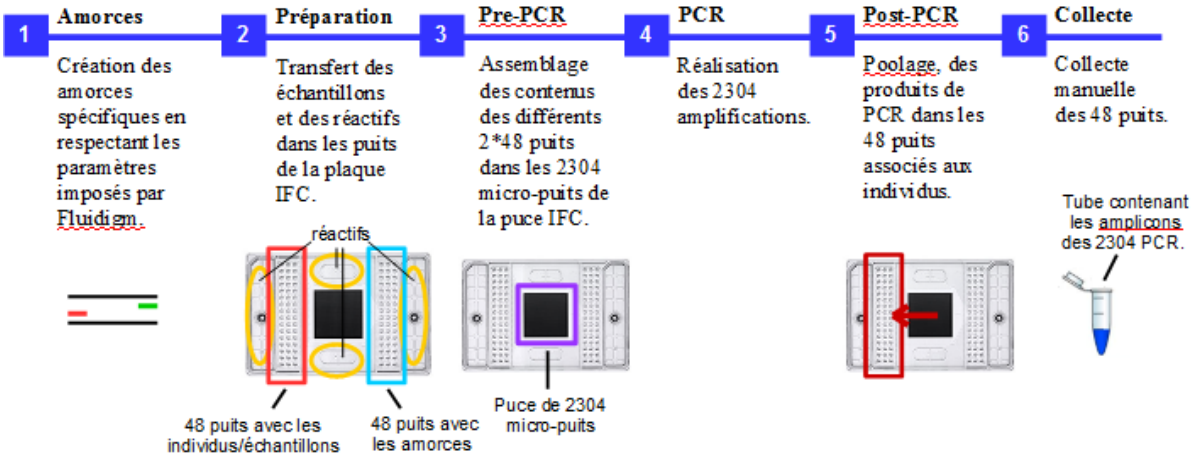


Figure 13 : Les différentes étapes de la PCR Access Array de Fluidigm (2015), N. Jacquier [25]

1- La stratégie de la création des amorces (partie 3.1)

2- Préparation de la plaque IFC par le manipulateur :

La plaque IFC se compose de 48 puits à gauche (rectangle rouge sur le schéma), où sont répartis par le manipulateur l'ADN des 48 individus étudiés [25]. A droite, se trouvent 48 puits où sont réparties les amorces (schéma rectangle bleu). Les réactifs (voir [26]) sont placés dans 6 puits différents (ovales jaunes).

3- La **pré-PCR** ou préparation de la plaque dans l'IFC Controller AX de Pré-PCR:

Au centre de la plaque se situe une puce contenant les 2304 micro-puits qui sont reliés aux 96 puits (individus et amorces). Un système de valves et de pression permet de mélanger le contenu des puits pour permettre les 2304 réactions d'amplification.

4- Déroulement de la **PCR** dans le FC1 Cycler :

La PCR Access Array se compose de 35 cycles [26], chacun constitué de trois grandes étapes (Fig.14), (Annexe VI) :

- (a) Les deux brins de l'ADN sont dénaturés, c'est-à-dire qu'ils sont séparés. Pour cela, la température est élevée à 95°C.
- (b) Dans un second temps, la température est diminuée pour permettre aux amorces de s'hybrider aux brins d'ADN. Cette température est variable selon les amorces. Cependant, les différentes températures étant standardisées, il est nécessaire que les 48 couples d'amorces désignés respectent tous une gamme de température définie. Ceci sera un paramètre important à prendre en compte lors de notre stratégie de création des amorces.
- (c) Il y a enfin une phase d'élongation qui consiste en la synthèse du brin complémentaire par une ADN-Polymérase.

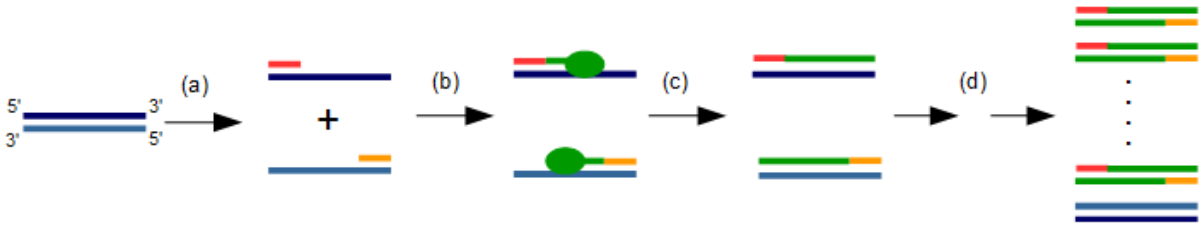


Figure 14 : Les différentes étapes d'une réaction de polymérisation en chaîne (PCR) *(2015) N. Jacquier*

5- La **post-PCR** dans l'IFC Controller AX de Post-PCR:

Les 48 amplicons issus des 48 couples d'amorces sont regroupés ensemble par individu.

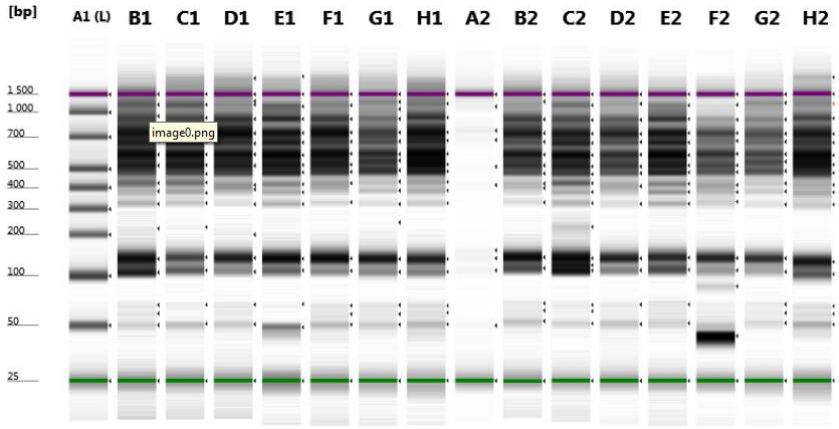
6- La collecte :

A l'issue de la post-PCR, un poolage manuel de tous les individus est réalisé. Il s'en suit une étape de purification des amplicons pour éliminer les primers-dimers (sous-produits de PCR issus de l'hybridation de deux amorces entre-elles) et les amorces restantes. C'est le résultat de ces étapes qui constituera la librairie qui sera séquencée par la suite.

Vérification du « run test » et du fonctionnement de l'amplification

Dans le cas du « run test », les résultats de 15 des 48 puits issus du produit non purifié de 48 PCR (Après le poolage par individu (étape 5) mais avant le poolage final (étape 6)) ont été testés sur gel High Sensitivity D1000 ScreenTape®. Ce test permet de savoir si la PCR a fonctionné de façon générale pour les individus mais pas de façon spécifique pour chaque couple d'amorces. En effet, seul le séquençage de ces amplicons dans le MiSeq permet de savoir plus précisément si les amplifications ont ou non fonctionné.

Ce sont ces données qui nous ont permis de définir le seuil de quantité minimale d'ADN nécessaire pour nos échantillons. Les résultats sur gel (Fig.15) permettent de visualiser la longueur des amplicons issus de la PCR. Plus la bande est foncée, plus il y a d'amplicons de la longueur correspondante. Dans le cas du « run test », on observe donc que la PCR a bien fonctionnée hormis pour l'échantillon A2 dont la concentration était bien trop faible. On peut également constater des bandes à 25pb qui correspondent aux amorces en surplus, et entre 50 et 130pb qui sont dues aux primer-dimers. Ceci est dû au fait que les ADN après l'amplification n'ont pas été purifiés.



Well	Conc. [pg/µl]	Sample Description
A1	2610	Ladder
B1	4370	B00G3SV-2ng/uL
C1	4520	B00G3SV-10ng/uL
D1	3700	B00G3SW-10ng/uL
E1	5440	B00G3SX-10ng/uL
F1	4350	B00G3T4-10ng/uL
G1	3960	B00G3T5-10ng/uL
H1	5290	B00G3T6-10ng/uL
A2	138	B00G3ST-5ng/uL
B2	3750	B00G3SU-5ng/uL
C2	5510	B00G3SV-5ng/uL
D2	3040	B00G3SW-5ng/uL
E2	4990	B00G3SX-5ng/uL
F2	4380	B00G3T4-5ng/uL
G2	3290	B00G3T5-5ng/uL
H2	4430	B00G3T6-5ng/uL

Figure 15 : Résultats de 15 produits de PCR non purifiés sur gel High Sensitivity D1000 ScreenTape®. *Le A1 correspond à un ladder qui sert de référence et permet la lecture par comparaison des longueurs des amplicons. Le tableau à gauche indique les échantillons de chaque puits ainsi que leur concentration à l'issue de la PCR et le code-barre des individus correspondants.*

Au vu de ces données « test », un seuil minimal de concentration en ADN de nos échantillons a été fixé de façon à obtenir une amplification générant suffisamment d'amplicons pour le séquençage ultérieur et donc des données exploitables. Cette valeur est placée à 7,5 ng/µL pour un dosage au fluorimètre. En effet, l'ADN du rosier A2, dont la concentration avait été mesurée à 5ng/µL, n'ayant pas été amplifié, une concentration supérieure a été choisie.

De plus, on constate des écarts entre les dosages des différents appareils. En effet, nous avons réalisé un dosage par fluorimétrie et l'EPGV à l'aide du Qubit®. Or, il n'y a pas de corrélation observée entre les deux mesures. Ainsi, pour plus de sûreté, seuls nos échantillons dont le dosage par fluorimétrie est au-dessus de 10ng/µL ont été envoyés. Cela a donc nécessité une réextraction de l'ADN de 192 de nos individus.

Le séquençage MiSeq d'Illumina®

A l'issue de la PCR et des étapes de contrôle et purification, un séquençage à haut débit, ou séquençage nouvelle génération (NGS), a été réalisé à l'aide du séquenceur MiSeq d'Illumina® (Fig.16) [27]. Cette étape permet à partir des amplicons purifiés de la PCR, de former des clusters, c'est-à-dire des groupes d'amplicons. Ceux-ci vont être « lus » par le séquenceur pour obtenir les données brutes en sortie qui sont appelées des reads.

Figure 16 : Séquenceur Illumina® MiSeq *[27]*

Cette technologie a été choisie en raison de la longueur des reads que l'on souhaitait obtenir, ainsi que pour la qualité et la profondeur (~nombre de reads/amplicon) des données. De plus, le coût de la production de ces dernières a imposé certaines limites quant au choix du nombre de gènes et d'individus analysés et des technologies employées.

Les différentes étapes du séquençage sont les suivantes :

1- Préparation du Flowcell (Fig.17) :

Tout d'abord, les amplicons de la PCR sont répartis sur un support solide en verre, appelé flowcell, où 2 types de fourches/adapteurs (Forward et Reverse) sont fixées une ligne, elle-même divisée en 14 sections appelées « tiles »[28].

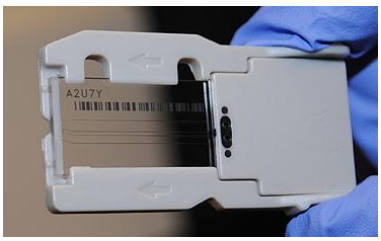


Figure 17 : Flowcell Illumina® MiSeq *(2015) [29]*

2- La formation d'un cluster (Fig.18):

Les amplicons, qui constituent notre librairie, vont s'hybrider aux fourches par complémentarité grâce aux adaptateurs NGS (Fig.18, adaptateurs en jaunes et verts). Le brin complémentaire va alors être synthétisé à la

suite des fourches et donc fixé au support par leur biais (a). Il vient ensuite la formation de clusters par amplifications en pont successives (de b à h). Après dénaturation, les brins reverse sont clivés et éliminés : c'est la linéarisation (i). Le séquençage du brin forward peut débuter (j).

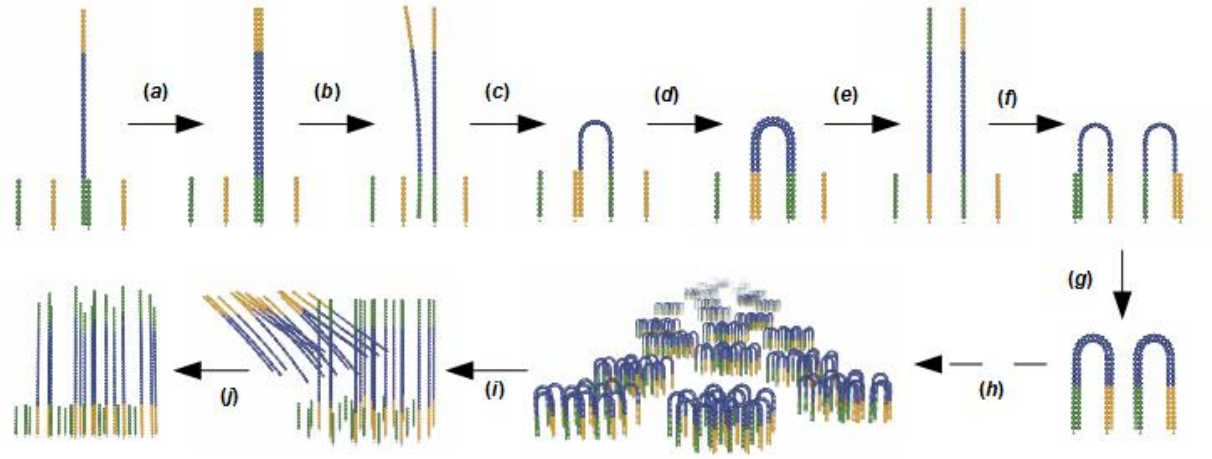


Figure 18 : Création de clusters par la technologie Illumina®. *Les différentes étapes sont : Fixation de l'amplicon par hybridation sur les fourches; (a) synthèse du brin complémentaire ; (b) dénaturation et élimination du brin d'origine ; (c) hybridation de l'extrémité du brin synthétisé sur une nouvelle fourche ; (d) Synthèse du brin complémentaire ; (e) Dénaturation ; (f) – (h) répétitions des étapes précédentes menant à la formation d'un cluster ; (i) Linéarisation ; (j) Début du séquençage par synthèse. Images sources [30]. (2015) N. Jacquier*

3- Réalisation du séquençage (Fig.19) [31] :

Il s'agit d'un séquençage à haut débit par synthèse, c'est-à-dire que la lecture simultanée des millions de brins par le séquenceur se fait au fur et à mesure de la fixation des ddNTP ou didésoxyribonucléotides. Ces derniers bloquent l'ajout de nouvelles bases et sont couplés à un fluorochrome dont la couleur est différente pour chacune des 4 bases (A, T, G ou C (1)). A chaque cycle de séquençage, il y a incorporation du ddNTP complémentaire de la base du brin à séquencer. Après sa fixation (2), le fluorochrome est excité et une fluorescence est émise (3). Celle-ci est alors reconnue/ « lue » par le séquenceur. Le bloqueur du ddNPT et le fluorochrome sont ensuite clivés (3) pour permettre la fixation du ddNPT suivant. (4) L'opération est renouvelée pour les 299 bases suivantes car pour notre étude, le nombre de cycles se porte à 300.

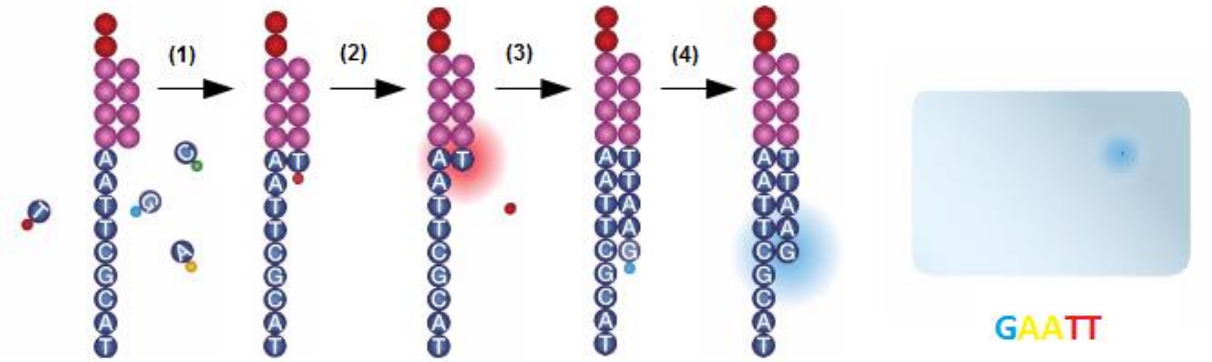


Figure 19 : Schéma du séquençage à haut débit. *Fixation de l'amorce de séquençage puis départ du premier cycle avec libération et compétition entre les ddNTP; (1) Incorporation du ddNTP complémentaire ; (2) excitation et émission d'une fluorescence caractéristique suivi d'un clivage du fluorochrome et du bloqueur; (3) (4) Ajout de nouveaux nucléotides. A droite illustration de la lecture par le séquençeur des bases d'un read. Source images [30]. (2015) N. Jacquier*

La lecture de milliers de reads est effectuée en parallèle. On obtient alors plusieurs centaines de fichiers avec les séquences des reads. Ce sont ces derniers qui feront l'objet de notre traitement bioinformatique.

Partie 3 : Résultats

Les résultats obtenus au cours de cette étude sont des stratégies qui concernent d'une part l'amont du projet avec la création des amorces et d'autre part le traitement des données issues du séquençage. Nous allons donc présenter dans un premier temps la méthode qui a été élaborée pour le design des amorces. Puis nous développerons les stratégies et les choix qui ont été mis en place pour analyser les données brutes issues du MiSeq et obtenir des fichiers de sortie permettant l'analyse de la diversité des gènes candidats. Nous verrons ainsi comment a été conçu notre workflow et quels outils ont dû être créés par la suite pour permettre de réaliser le contigage et le dosage des haplotypes.

3.1. Annotation des gènes candidats et mise en place d'une stratégie de conception d'amorces

Comme nous l'avons expliqué brièvement précédemment, il est nécessaire d'annoter et définir les différentes zones de nos séquences de référence avant de créer nos amorces. Nous avons vu en partie 2 quelle méthode était appliquée pour trouver nos séquences sources et nous allons donc maintenant expliquer comment sont annotées ces séquences et expliquer l'importance de cette étape qui est déterminante pour la création de nos couples d'amorces. Nous pourrons alors expliquer la stratégie qui a été adoptée pour permettre d'optimiser le choix de nos amorces et répondre aux mieux à nos objectifs tout en prenant en compte de nombreuses contraintes.

3.1.1. L'annotation des séquences de référence des gènes candidats

Après avoir sélectionné les séquences sources, plusieurs opérations sont réalisées à l'aide du logiciel Geneious® pour annoter les séquences de référence. Ces étapes sont les suivantes :

- Il faut tout d'abord réaliser un « Blastn » [32] entre la séquence source annotée préalablement choisie et le génome du rosier (Altschul et al., 1990). Le Blastn permet de comparer la séquence source et les séquences des scaffolds de rosier. Il liste les différents scaffolds qui présentent la meilleure similarité de séquence avec la séquence source. On fait l'hypothèse que les séquences présentant une forte similarité correspondent à un orthologue du gène de référence chez le rosier.
- Parmi les résultats des scaffolds signalés à l'issue du blast, seuls 2 sont retenus. En effet, il se trouve que certains fragments du gène peuvent blaster avec de nombreux scaffolds en raison d'une succession de bases identiques à d'autres portions du génome comme c'est le cas chez différents gènes des voies de biosynthèse. Cependant, la ploïdie de 'Old Blush' (2n=2x, d'où au maximum deux allèles) et sa forte hétérozygotie liée à son origine hybride (généralement deux allèles par locus) implique la sélection de 2 scaffolds, correspondants à 2 allèles. Ceux retenus sont ceux ayant la e-valeur la plus faible et dont le pourcentage de séquences blastées est le plus élevé possible. La e-value traduit le degré de certitude d'homologie entre les deux séquences. Ainsi, plus la e-value est faible (inférieure à 10^{-10}), plus les chances d'avoir un bon blast sont élevés.
- Les séquences des deux scaffolds sont alors récupérées.
- Les exons de la « séquence source », et les extrémités 5' et 3'UTR lorsqu'elles sont disponibles, sont extraits individuellement (Fig.20, bleu).
- Un assemblage [32] entre les scaffolds, les exons et les extrémités est effectué (Fig.20, A). Il s'agit d'un alignement local optimisé permettant de détecter des similarités entre des petites séquences chevauchantes ou non et une plus grande séquence.

- Les exons et les extrémités UTR des scafflods du rosier sont alors repérées et annotés, puis l'annotation des introns est ajoutée (Fig.20, C).

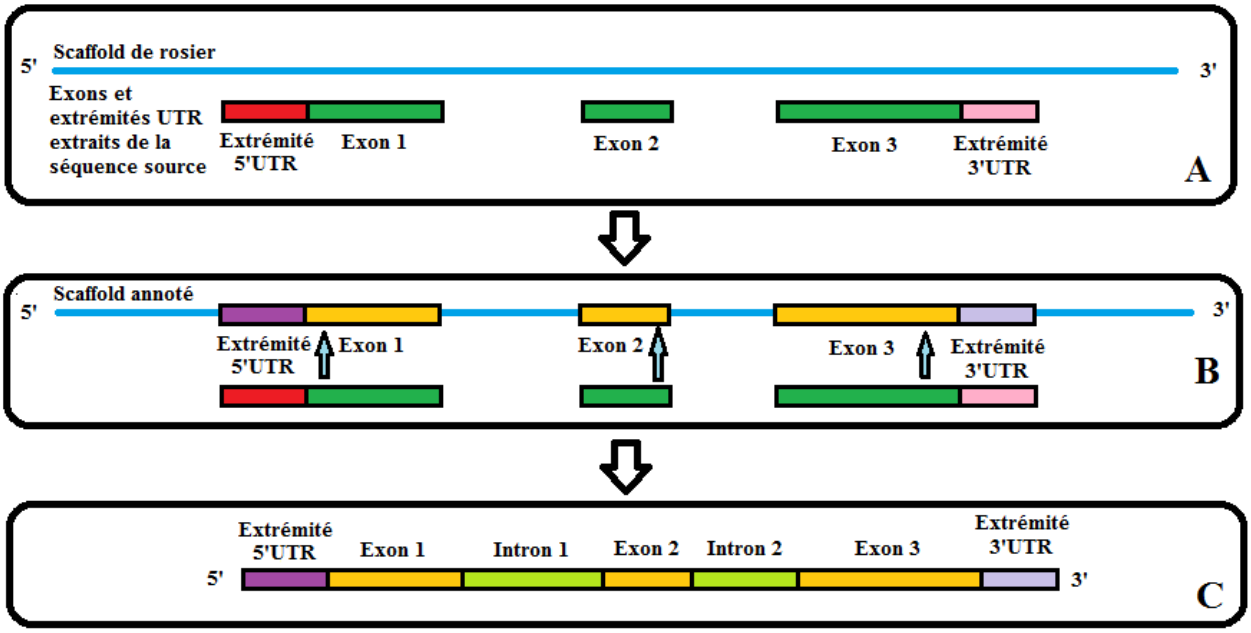


Figure 20 : Schéma d'annotation des gènes candidats du rosier. *En A : Assemblage du scaffold du rosier et des exons et extrémités UTR ; B : Comparaison des séquences et annotation du scaffold ; C : Annotation des introns. (2015). N.Jacquier*

3.1.2. La recherche de polymorphismes et des zones d'intérêt

A l'issue de l'annotation des gènes, il est important d'avoir un maximum d'informations sur leur polymorphisme. En effet, au sein des gènes, les introns sont moins bien conservés que les exons et présentent plus de mutations. Les amorces à créer sont donc placées dans les exons. Cependant, là encore il peut y avoir des variations de bases entre ou au sein des individus, appelées SNP ou Single Nucleotide Polymorphism.

Ces SNP sont à double tranchant. Certains peuvent avoir un intérêt pour leur importance fonctionnelle car ils expliquent des différences importantes d'un point de vue biologique et phénotypique comme pour le gène *DFR* (partie 1). De plus, certains SNP sont parfois caractéristiques d'un allèle donné et favorisent son identification. Dans ces cas-là, il faut prendre soin de bien positionner les couples d'amorces pour veiller à ce que ces SNP d'intérêt soient séquencés pour notre étude et qu'ils ne soient donc pas positionnés au sein des amorces. Mais d'un autre côté, lorsque l'on ne connait pas la présence et la position de telles mutations, cela peut poser problème pour la fixation des amorces ou bien entraîner l'amplification préférentielle d'un allèle par rapport à un autre. Ceci est particulièrement gênant car cela peut biaiser le dosage d'haplotypes. Enfin, les SNP peuvent parfois être utilisés pour empêcher la fixation des amorces sur des gènes proches dans le cas de familles multigéniques ou de gènes ayant des portions de leur séquence similaires. C'est par exemple le cas pour les couples d'amorces du gène *FT* qui devaient permettre d'amplifier *FT* mais pas *FT3* ou bien des couples d'amorces du gène de *NUDX1-OB1* qui ne devaient pas amplifier *NUDX1-OB2*.

C'est pour toutes ces raisons qu'il est important de récupérer autant que possible des informations quant à la position de ces SNP. Les deux scaffolds annotés de 'Old Blush' apportent déjà une première indication du polymorphisme du gène en question. Mais, lorsque cela est possible, des séquences de transcrits de différents rosiers sont récupérées et alignées avec les exons des gènes de rosiers pour identifier la position de SNP, d'insertions ou de délétions (Fig.21).

3.1. Annotation des gènes candidats et mise en place d'une stratégie de conception d'amorces

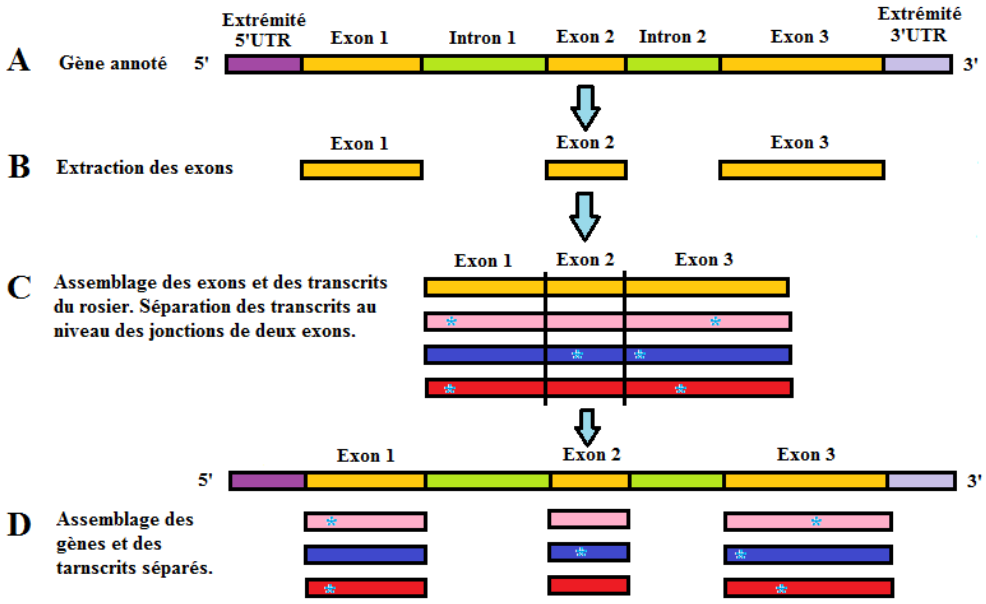


Figure 21: Identification du polymorphisme au sein d'un gène candidat. ** indique la présence de SNP ou Indels. A : Récupération de la séquence annoté du gène candidat chez le rosier ; B : Extraction des exons du gène ; C : Alignement des exons et des transcrits du gène chez le rosier ; D : Assemblage des exons et des transcrits. (2015). N. Jacquier*

Bilan des informations des gènes candidats à l'issu de l'annotation

Le tableau ci-dessous (Tab.3) récapitule les informations de la structure de chaque gène candidat chez le rosier.

Tableau 3 : Tableau récapitulatif de la structure de chaque gène candidat. *Informations sur la structure du gène, les numéros et longueurs des scaffolds où ils se situent. (2015). N. Jacquier*

Gène	Nombre d'introns	Nombre d'exons	Présence 5'UTR	Présence 3'UTR	N° du scaffold	Longueur (pb)
DFR	5	6	oui + Intron 5'UTR	Oui	2609	2309
					5050	2286
F3'H	2	3	Non	Non	130	2477
					2731	2607
FLS	2	3	Oui	Oui	3042	5529
					3192	4515
ZDS	13	14	Oui	Oui	1229	4819
					3026	4807
AGAMOUS	7	8	oui + Intron 5'UTR	Oui	161	6535
					1924	6585
FT	3	4	Non	Oui	4318	1238
					3636	1315
KSN	3	4	Non	Non	5123	1186
					R. chinensis spontana - Transposon	10114
					R. chinensis spontana	1184
NUDX	1	2	Non	Non	OB1a	622
					OB1b	618
					Papa meilland	618
					Blunt	618
					1636	623
					5234	623
					866	606
PAAS	0	1	Non	Non	1004	1526
					1103	1527
BRC1	1	2	Non	Non	1427	1555
					1951	1544

3.1.3. La stratégie de création des amorces

Objectif général de la stratégie

Pour réaliser les amorces c'est le logiciel Primer3 v.0.4.0©, inclus dans Geneious Pro 5.3.6© qui a été utilisé (Rozen & Skaletskt, 2000). Les amorces ont été commandées par l'EPGV.

L'objectif général de notre stratégie de design des amorces est de séquencer l'ensemble du gène candidat tout en minimisant le nombre de couples nécessaires. Plusieurs paramètres entrent en compte lors de la création des amorces de sorte que parfois le choix de celles-ci s'est révélé très restreint, voire unique. Notre stratégie a donc consisté en la réalisation d'un compromis entre nos objectifs d'amplification et l'optimisation du choix des amorces malgré les nombreuses contraintes imposées par la technique Access Array et la nature de la séquence d'ADN elle-même.

Les contraintes prises en compte dans la stratégie de design des amorces

Il y a différents types de contraintes à considérer :

- Des contraintes techniques liées aux paramètres de la PCR et du séquençage MiSeq.
- Des contraintes biologiques dues à la diversité des séquences d'ADN, la présence de SNP, la longueur des exons…
- Des contraintes liées à nos objectifs tels que le séquençage de l'ensemble du gène, le nombre réduit de couples d'amorces à utiliser…
- Des contraintes liées à l'analyse bioinformatique des données de séquençage comme le besoin d'une zone de chevauchement intra et inter-séquence pour les phases de contigage et de reconstitution des haplotypes.

Nous allons maintenant expliciter ces contraintes pour pouvoir par la suite les hiérarchiser. La première d'entre elles est due au fait que les amorces doivent être placées dans les exons qui sont mieux conservés que les introns et donc moins sujets aux SNP. En effet, un SNP peut fortement diminuer la fixation des amorces sur le brin d'ADN ou amplifier un allèle de façon préférentielle. Cela influence donc sur la longueur des amplicons situés entre les couples d'amorces que l'on va mettre au point. Néanmoins, dans certains cas des amorces dans les introns sont réalisées car :

- la longueur entre deux exons est trop importante
- l'intron est très long et doit être séquencé en plusieurs fois (ex : gène *FLS*, Fig.24)
- aucune amorce n'est possible dans les exons du fait des contraintes de composition des amorces et/ou de la présence de polymorphisme dans la région correspondante.

Compte tenu des contraintes imposées par Fluidigm®, la plateforme de séquençage d'Evry nous a communiqué certains critères à respecter dont des températures de fusion « Tm » (melting temperature) et le nombre maximum de Poly-X toléré dans l'amorce « Max Poly-X ». Celui-ci correspond au nombre maximal de succession d'une même base au sein de l'amorce. Comme nous l'avons vu dans le protocole de la PCR Access Array, les Tm sont standardisées durant les cycles de PCR et doivent donc être dans une certaine fourchette puisque 48 couples d'amorces sont amplifiés en même temps et ont donc tous les mêmes conditions.

Les fourchettes de ces deux paramètres sont modifiables dans une certaine limite mais de façon plus réduite que les paramètres par défaut de Primer3.

Pour Fluidigm® les critères en rouge (Annexe VII) sont conseillés d'être :

- Entre 59 et 61°C pour la Tm
- Max Poly-X=3

3.1. Annotation des gènes candidats et mise en place d'une stratégie de conception d'amorces

La technologie du séquençage MiSeq utilisée dans le cadre de cette étude permet d'avoir des longueurs de reads pouvant atteindre 300 bases. Ainsi, le read forward et le read reverse contigués feront au maximum 600 bases. Les amplicons à produire ne devront donc pas, dans la limite du possible, excéder 600pb. Cependant, certains sont de taille supérieure car : il n'y a pas d'autres choix (pas d'amorces plus proches entre-elles) ou bien cela permet de récupérer des zones non couvertes par d'autres couples d'amorces (Fig. 24, gènes *ZDS* et *AGAMOUS*).

De plus, il est important de conserver un chevauchement intra-séquences de 50 pb pour le traitement bioinformatique des données (Fig.22). Il est connu que la qualité de la séquence des reads diminue fortement en fin de séquence. Cela permet d'autre part d'avoir une zone de chevauchement entre les reads forward et reverse et donc de réaliser un contigage intra-séquence, c'est-à-dire de rassembler les 2 reads en un seul. C'est pourquoi il a été privilégié une distance entre les amorces d'un même couple de 550pb (ou moins)

A cela s'ajoute la nécessité de conserver si possible un chevauchement inter-séquences supérieur à 50 pb (Fig.22). Cela permettra de rassembler les différents fragments séquencés entre eux après la phase de dosage des haplotypes, afin d'obtenir des haplotypes plus longs.

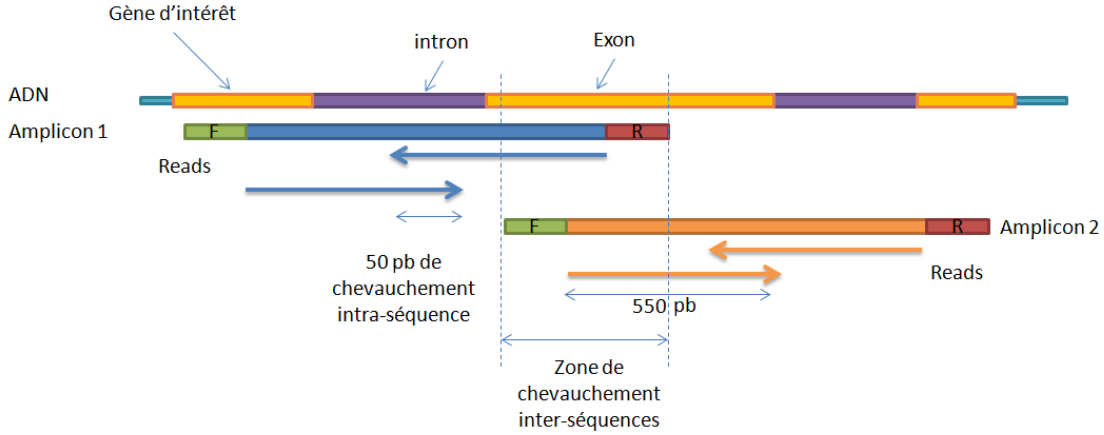


Figure 22 : Schéma de constitution des amorces. *Indication des zones de chevauchement intra-séquence et inter-séquences. F indique une amorce Forward et R une Reverse (2015). Mathilde Liorzou.*

Pour finir, la spécificité des amorces a été vérifiée pour chaque couple après la phase de design. Cette étape permet de contrôler que les amorces n'amplifient pas d'autres zones du génome. Pour cela, il a été réalisé à l'aide de Geneious Pro 5.3.6© un « discontiguous Megablast » avec une e-valeur maximale de 0. Si, pour un couple d'amorce donné, des scaffolds autres que ceux comprenant le gène sont trouvés, on réalise une comparaison des résultats donnés pour l'amorce Reverse et pour l'amorce Forward :

- S'il y a des résultats communs, on regarde :
    - quel est l'intervalle entre les deux amorces sur ce nouveau scaffold. Si celui-ci est très grand, on suppose que la PCR ne fonctionnera pas sur le scaffold parasite.
    - quel est le pourcentage de séquence de l'amorce qui est identique à la zone blastée. Si les amorces ne sont que partiellement complémentaires (moins de la moitié de la longueur), elles ne permettront pas d'amplifications ou tout du moins pas d'amplification de la zone souhaitée et elles sont conservées.
    - ➔ Dans ce cas, on déplace les amorces et on revérifie leur spécificité. Ce cas de figure s'est présenté plusieurs fois notamment pour les gènes AGAMOUS (Entre des exons 2 à 6) et ZDS (entre les exons 6 et 7 et des exons 12 à 14).
- Sinon, la spécificité est bonne et nos amorces sont validées.

3.1. Annotation des gènes candidats et mise en place d'une stratégie de conception d'amorces

La hiérarchisation des critères pour la stratégie de design des amorces

L'ensemble de ces différents critères a été hiérarchisé du plus important au moins important de façon à optimiser les amorces et le traitement ultérieur des données.

1- Pas de SNP
2- Spécificité des amorces
3- Longueur de l'amplicon (optimum de 550pb comprenant le chevauchement intra-séquence) nécessaire permettant un contigage intra-séquence ultérieurement
4- Conservation d'un chevauchement inter-séquences (>50pb si possible)
5- La Tm et le Max Poly-X (optimisés puis, dans la limite des paramètres par défaut)

Certains de ces points sont étroitement liés puisque si le polymorphisme du gène n'est pas connu, il est possible d'avoir des SNP au sein des amorces sans le savoir et donc d'avoir des problèmes de spécificité. Générer des amorces dans certaines zones s'avère donc parfois long.

Une stratégie particulière adaptée pour le cas du gène KSN

Certains allèles de *KSN* présentent l'insertion d'un rétrotransposon d'environ 10 kb, situé dans l'intron 2. Pour pouvoir amplifier cette zone quel que soit l'allèle, une stratégie a été mise en place à partir de 2 couples d'amorces (ceux présentés en B, Fig.23) qui seront mis ensemble dans un même puits de PCR. En présence du rétrotransposon (B), deux amplicons seront obtenus. En son absence (A), l'amorce forward du premier couple d'amorce (en rouge) fonctionnera avec l'amorce reverse du second couple d'amorce (en magenta). Il y aura donc un seul amplicon produit. Dans cette stratégie, les hétérozygotes donneront trois amplicons différents.

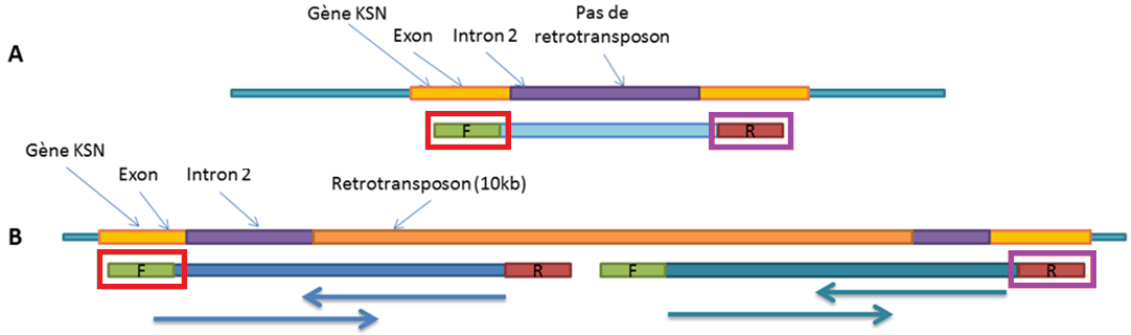


Figure 23: Stratégie de création des amorces pour le gène KSN. *2 couples d'amorces permettent d'amplifier les allèles avec ou sans rétrotransposon. Schéma non à l'échelle.(2015). Mathilde Liorzou.*

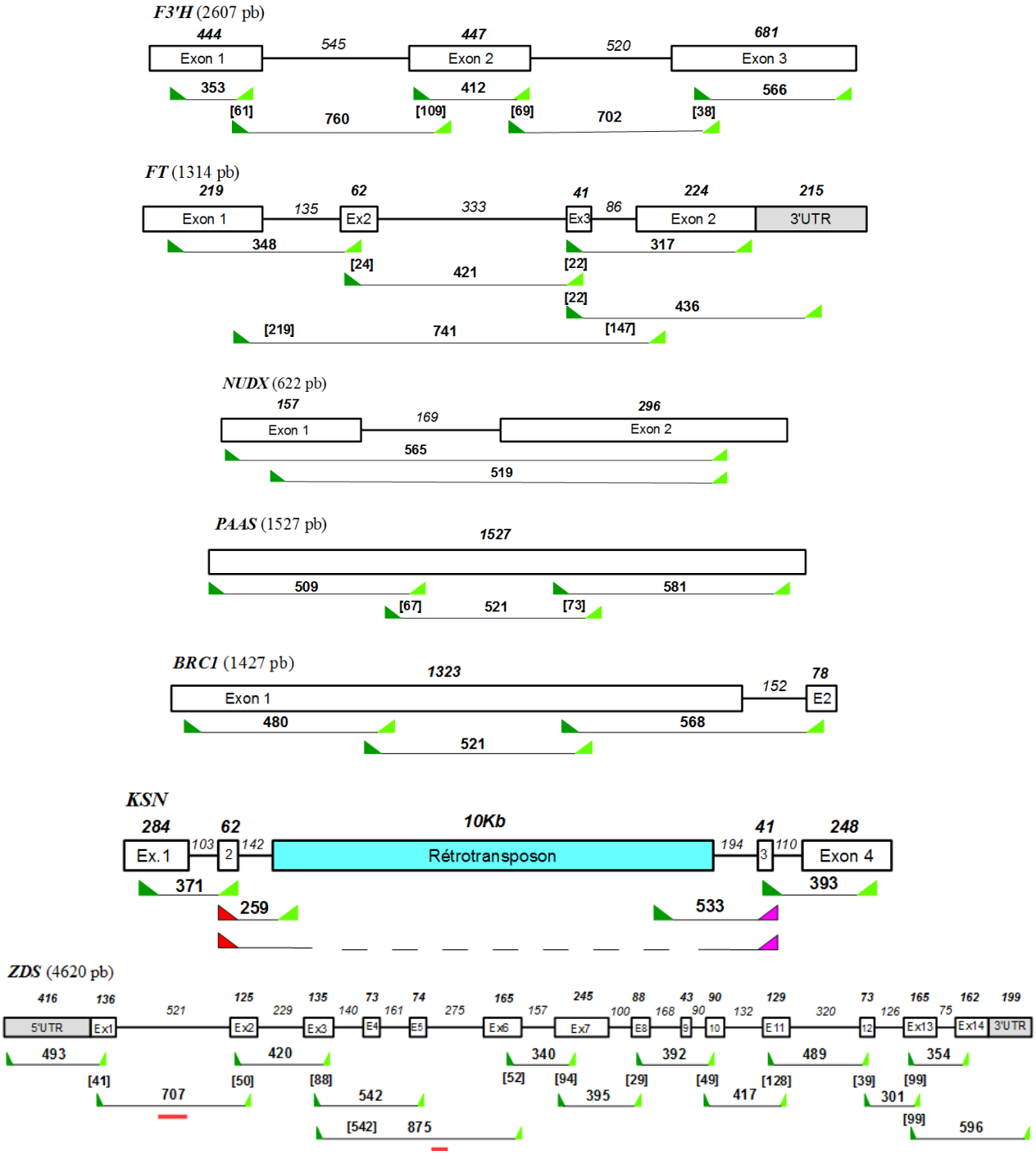
Bilan des amorces mise au point

Le tableau suivant (Tab.4) synthétise toutes les informations connues concernant les paramètres utilisés pour la création des amorces ainsi que la présence de SNP.

Tableau 4: Tableau récapitulatif des informations concernant les amorces. *NC= non connu en raison d'un scaffold contenant de nombreuses bases N, c'est-à-dire non identifiées. (2015) N. Jacquier*

Gène	Nombre de couples d'amorces	Nombre d'amorce avec...				Remarques
		1 SNP	Max Poly-X =4	Max Poly-X =5	Tm inférieure à 59°C	
DFR	6	3	2	0	5	
F3'H	5	0	0	0	2	
FLS	12	NC	9	1	2	14 amorces dans introns
ZDS	13	3	4	2	7	
AGAMOUS	14	0	4	1	12	12 amorces dans introns
FT	5	1	0	2	9	
KSN	4	0	0	0	4	
NUDX	2	2	0	0	2	
PAAS	3	0	2	0	2	
BRC1	3	1	3	0	3	

Ceci permettra de déterminer l'échec de certaines amorces à l'issue de l'amplification et du séquençage (Annexe VIII, informations pour chaque couple d'amorces). Les schémas ci-dessous (Fig.24) représentent les positons de l'ensemble des couples d'amorces des différents gènes candidats. Les chiffres entre crochets indiquent la zone de chevauchement inter-séquences, amorces comprises (en vert), et la longueur des amplicons est indiquée entre les deux amorces (en pb). Enfin, la taille (en pb) des introns et des exons est indiquée au-dessus de ceux-ci et les traits rouges signalent des zones qui ne seront pas séquencées.



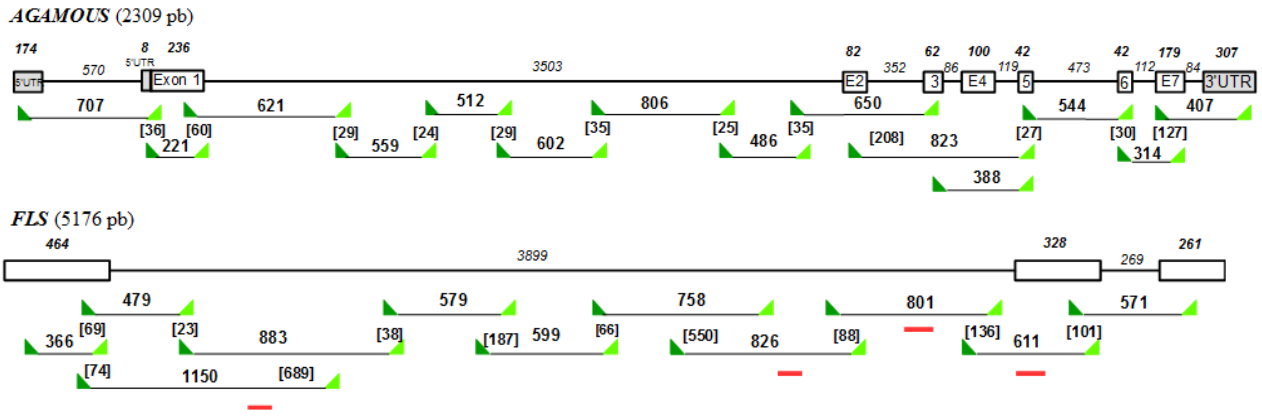


Figure 24 : Schémas bilans à l'échelle des couples d'amorces créés pour chaque gène candidat. *Les amorces sont indiquées en vert foncé pour les forward et en vert clair pour les reverse. Entre crochets est précisé la longueur de la zone de chevauchement inter-séquences. Entre les amorces est signalée la longueur de l'amplicon. Les exons sont représentés par des rectangles et les introns par des traits. Leur taille en pb est indiquée au-dessus. Les traits rouges signalent l'absence de données à l'issue du séquençage en raison d'une taille d'amplicon trop élevé où de zones non couvertes par des amorces. (2015). N. Jacquier*

3.2. Réalisation d'un workflow

Bien que le schéma général de traitement et d'analyse des données soit établi, il n'y a pas de règles prédéfinies et le choix des logiciels à utiliser reste au bon vouloir de l'utilisateur. Néanmoins, la quantité importante de données à traiter impose la réalisation automatique de différentes opérations. On met alors en place un workflow, c'est-à-dire une suite d'opérations qui s'enchaînent sans intervention extérieure à partir de fichiers d'entrée. Il est donc nécessaire de définir les données que l'on a en amont du workflow et celles que l'on souhaite obtenir en aval pour mettre en place le schéma le mieux adapté à nos besoins.

Nous verrons donc quel est le schéma général du workflow qui a été mis en place et quelles données sont disponibles (schéma complet et paramètres fixés en Annexe IX). Nous expliquerons aussi comment ont été choisis les logiciels ainsi que leurs paramètres, que ce soit à partir de données bibliographiques ou de test. Enfin, nous verrons les données de sortie obtenues.

3.2.1. Explication générale du workflow

Dans le cadre de notre étude, nous avons choisi l'instance Galaxy ABiMS de Roscoff qui met à disposition des utilisateurs de nombreux programmes et permet la mise en place d'un workflow dont le schéma général est le suivant (Fig.25):

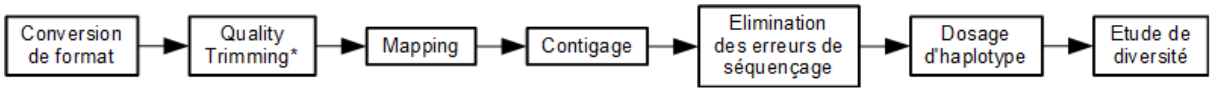


Figure 25 : Schéma général du workflow. *(*= élagage des reads par la qualité) (2015). N.Jacquier*

- Conversion de format (**FASTQ Groomer** version 1.0.4) [33] : Etape préalable au traitement des données qui permet une compatibilité avec les différents programmes.
- Quality Trimming (**FASTQ Quality Trimmer** version 1.0.0) [33]: Elagage des reads par élimination des bases de mauvaise qualité.
- Mapping (**Bowtie2** version 0.2) [34]: Alignement des reads contre la séquence de référence des gènes candidats. Cette étape est suivie d'une conversion de format (BAM-to-SAM (version, 1.0.4) [35]).
- Contigage : Assemblage des reads forward et reverse.
- Elimination des erreurs de séquençage

- Dosage d'haplotypes : Comptage des différents allèles des gènes candidats pour chaque individu.

3.2.2. Les données issues du MiSeq

A l'issue du séquençage, on obtient 2 fichiers par individus d'extension .fastq comprenant les reads 1 (Forward) dans le premier et les reads 2 (Reverse) dans le second. Dans chacun de ces fichiers, les reads issus des amplicons des 48 couples d'amorces sont mélangés et sont environ au nombre de 48 000 reads soit environ 1000 reads par couples d'amorces. Ce nombre est variable selon les individus (48087 pour 'Joséphine Ritter', 45393 pour 'Georges Delbard' par exemple).

Les fichiers de sortie (Fig.26 et Annexe X) indiquent un identifiant unique pour chaque read ainsi que la séquence nucléotidique du read et la qualité de chaque base [36]. De plus, les reads forward et reverse ont le même identifiant hormis un chiffre (carré rouge) qui indique si le read est forward (1) ou reverse (2). Il est important de souligner la présence des amorces spécifiques qui, contrairement aux adaptateurs, n'ont pas été retirées des séquences de 300 pb des reads.

Identifiant du read	@M01075:79:000000000-ACRPA:1:1101:17353:1093 2:N:0:73
Séquence du read	NGTGGCTTCGGAAGCTTGTCCCACTCATACCATCCCCAACCATCACAGAATTCTGGCCC
Identifiant de la qualité	+
Séquence qualité	#-8B@C,6E@6@,+,C;ECCF<FF,C@FFFG@<,CFGCEE7CAF9<F@8FC,CC,6E8F

Figure 26 : Extrait de fichier de sortie du MiSeq. *Il s'agit du read 2 de l'accession 'Joséphine Ritter' (2015). N. Jacquier*

Il faut également vérifier l'encodage des données afin de choisir les bons modes de conversion. En effet, les valeurs de qualité ont différentes échelles et se traduisent par différents caractères. Nos données sont encodées en Illumina 1.9, ce qui se traduit par la présence caractéristique de certains symboles [36] tels que : `!"#$%&'()*+,-./0123456789`. Après avoir identifié le codage utilisé (Annexe XI), une conversion de format [33] a été appliquée à l'ensemble des fichiers .fastq pour obtenir un codage Sanger.

3.2.3. Le quality trimming

Principe

Au sein d'un même read, toutes les bases n'ont pas la même valeur de qualité, c'est-à-dire que la probabilité que la base soit correcte n'est pas la même. Ainsi, plus la valeur du score de qualité d'une base donnée est élevée, plus la probabilité d'avoir la bonne base est forte. De façon générale, la qualité des bases diminue vers la fin du read, à l'extrémité 3' [37]. A noter que les reads 1 ont une meilleure qualité que les reads 2, sans qu'une explication à cela n'ait été déterminée.

Il est donc nécessaire de réaliser un filtrage des reads en éliminant les bases les plus mauvaises de façon à avoir des données de séquençage contenant le moins d'erreurs et d'incertitudes possibles.

La qualité de nos données

Pour visualiser la qualité de nos reads, nous avons utilisé le logiciel **FastQC** (version 0.62), intégré à Galaxy [37]. Il représente la qualité des bases du read de façon visuelle, ce qui permet une analyse rapide de l'effet des paramètres choisis pour notre trimming. Sur les graphiques de données (cf ci-dessous), des boîtes à moustaches sont créées pour chaque position de base et 3 zones distinctes peuvent être observées :

- En vert, les bases de bonne et très bonne qualité (qualité > 28)
- En orange les bases de qualité moyenne (20 < qualité < 28)

- En rouge les bases de mauvaise qualité (qualité < 20)

Sur cet exemple chez le rosier 'Georges Delbard', on observe nettement la diminution de la qualité avec la longueur du read, plus particulièrement au niveau des 100 dernières bases. Après une comparaison des qualités entre différents individus (2 diploïdes, 2 triploïdes, 2 tétraploïdes et 2 hexaploïdes), il s'avère que l'ensemble des profils sont très similaires (Fig.27) et que la zone à trimmer concerne essentiellement les 100 dernières bases.

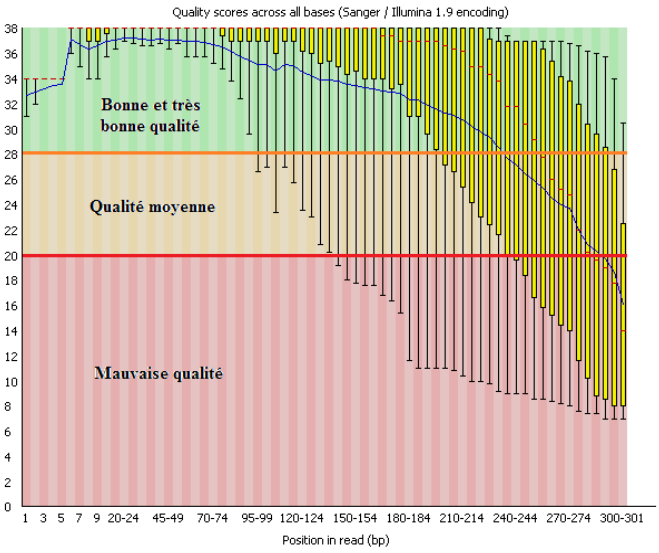


Figure 27 : Exemple de résultat des scores de qualité des bases du read 2 de l'accession 'Georges Delbard.' *Réalisé sur FastQC avant quality trimming. (2015). N. Jacquier*

Objectifs généraux du quality trimming adapté à nos données de séquençage

Le quality trimming est un compromis entre la conservation de toutes les bases des reads de séquençage et l'élimination d'un trop grand nombre d'entre-elles, entraînant une perte d'informations. Les objectifs pour parvenir à cela sont :

- Faire que la moyenne et la médiane des scores de qualité soit supérieure à 28 (zone verte) et qu'au moins 90% des scores des bases soit au-dessus de 20% (zone orange).
- Eliminer des zones de mauvaise qualité le plus loin possible au sein du read.

Les reads 2 étant de moins bonne qualité, c'est sur eux que les différents paramètres ont été testés. En effet, pour les reads de bonne qualité, même si les paramètres sont restrictifs, le trimming s'arrêtera de lui-même plus tôt.

Choix de la méthode de trimming

Plusieurs outils permettent de trimmer à partir de l'extrémité 3' du read dont le Fastq quality trimmer. Ce programme permet de choisir entre deux techniques de trimming. La première technique est dite de « fenêtre glissante ». Elle consiste en un calcul de moyenne ou somme de valeurs de qualité de plusieurs bases adjacentes. Si le résultat ne respecte pas une certaine condition, alors les bases sont éliminées. La seconde consiste au retrait des bases à partir de l'extrémité 3' une par une lorsque leur qualité ne dépasse pas une valeur seuil. Cela correspond à une fenêtre glissante de taille 1. L'approche par la méthode de fenêtre glissante est jugée plus intéressante [37]. En effet, elle permet un nettoyage du read plus approfondi et non pas un arrêt précoce du trimming dès le premier score de qualité au-dessus de la valeur seuil rencontrée.

Paramètres choisis

La fenêtre glissante réalise l'élimination des bases de mauvaise qualité en partant de l'extrémité 3' par pas de 1. On réalise une moyenne des scores de qualité des bases présentes au sein de la fenêtre et ce résultat doit être supérieur à 25 pour que la suppression de base s'arrête. (Annexe IX). Comme nous le verrons par la suite, il est essentiel de cocher la case « garder les reads avec une longueur de 0 bases ».

Test de différente taille de fenêtres glissantes

Le schéma ci-dessous (Fig.28) représente le trimming d'un read en fonctions de différentes tailles de fenêtres glissantes et des scores de qualités de chaque base de la séquence du read. Cela permet de mieux visualiser le fait que des fenêtres de taille réduite risquent d'arrêter le trimming de façon plus prématurée que des fenêtres de grande taille. En effet, ces dernières permettent de dépasser des zones de bonne qualité alors que les bases suivantes sont moins bonnes (ex : Séquence GAT incluse dans la fenêtre de taille 3). Cependant, une fenêtre de taille trop importante n'est pas forcément utile car lorsque la valeur seuil est atteinte, la moitié des bases de la fenêtre est conservée mais peuvent avoir des scores de qualité médiocre. Il faut donc un juste milieu entre une fenêtre top étroite et trop large.

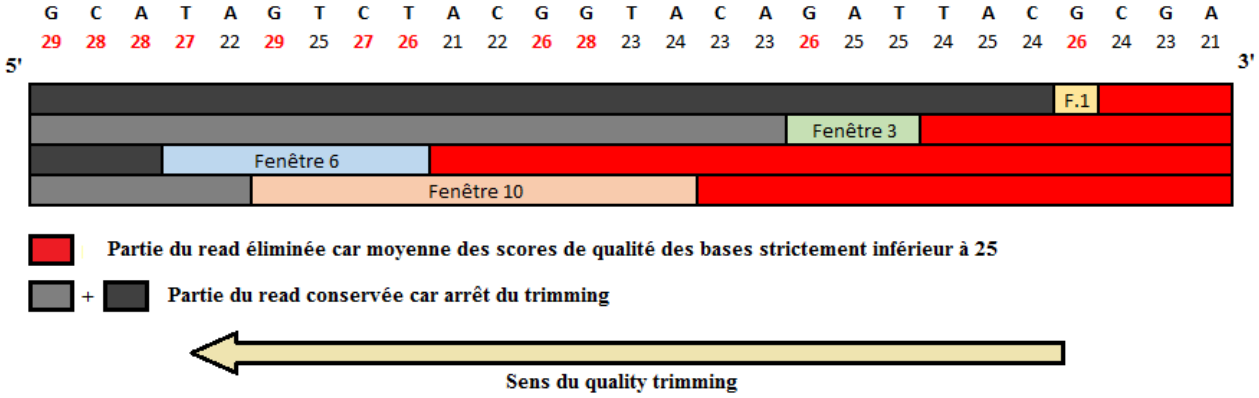


Figure 28 : Comparaison du quality trimming d'un read à l'aide de 4 fenêtres glissantes de différentes tailles pour une valeur seuil de 25. *La première ligne indique les bases de la séquence du read et au-dessous sont précisées la valeur de qualité de chaque base. Les zones rouges indiquent les parties du read éliminées, les zones grises (gris clair et foncé pour une meilleure lisibilité) indiquent les zones du read conservées car non affectée par le quality trimming. Enfin, les fenêtres sont représentée à l'endroit où elles se trouvent lorsque le quality trimming s'arrête. Les bases des fenêtres sont donc conservées dans la séquence du read final. (2015). N. Jacquier*

Choix d'une valeur seuil

Durant le trimming, il faut définir une valeur de qualité, que nous appellerons valeur seuil, pour laquelle les bases ayant un score inférieur sont éliminées. Cependant, il n'existe aucune indication quant à son choix qui reste donc à définir par l'utilisateur [60]. Néanmoins, Illumina indique que les bases des reads issus d'un séquençage MiSeq ont pour plus de 90% d'entre-elles un score de qualité supérieur à 30 [38].

Ainsi, de façon à répondre à nos objectifs, nous avons comparé les résultats de différents paramètres de trimming avec des valeurs seuil de 20, c'est-à-dire à la limite de la mauvaise qualité, et de 25, qualité moyenne. Une valeur supérieure à 25 ne semblait pas adaptée, car avec la technique de fenêtre glissante, une valeur de qualité trop élevée risquait d'éliminer beaucoup de bases au score correct (compris entre 24 et 28).

Test de différents paramètres

Nous avons donc comparé les résultats de quality trimming réalisé avec différentes taille de fenêtre et de seuils. Les paramètres fixés sont :

- la fenêtre glissante réalise l'élimination des bases de mauvaise qualité en partant de l'extrémité 3' par pas de 1.
- On réalise une moyenne des scores de qualité des bases présentes au sein de la fenêtre et ce résultat doit être supérieur à la valeur seuil pour l'arrêt du trimming

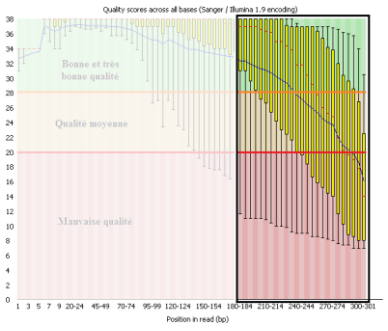


Figure 29 : La zone encadrée représente les dernières bases du read qui seront le plus affectée par le quality trimming. *(2015) N. Jacquier*

- La case « garder les reads avec une longueur de 0 bases » est cochée car cela sera important pour le mapping.

Les résultats présentés sont ceux du read 2 de l'accession 'Georges Delbard'. Pour des raisons de clarté, seule les dernières centaines de bases seront présentées (Fig.29, Fig.30), le début différant peu après traitement.

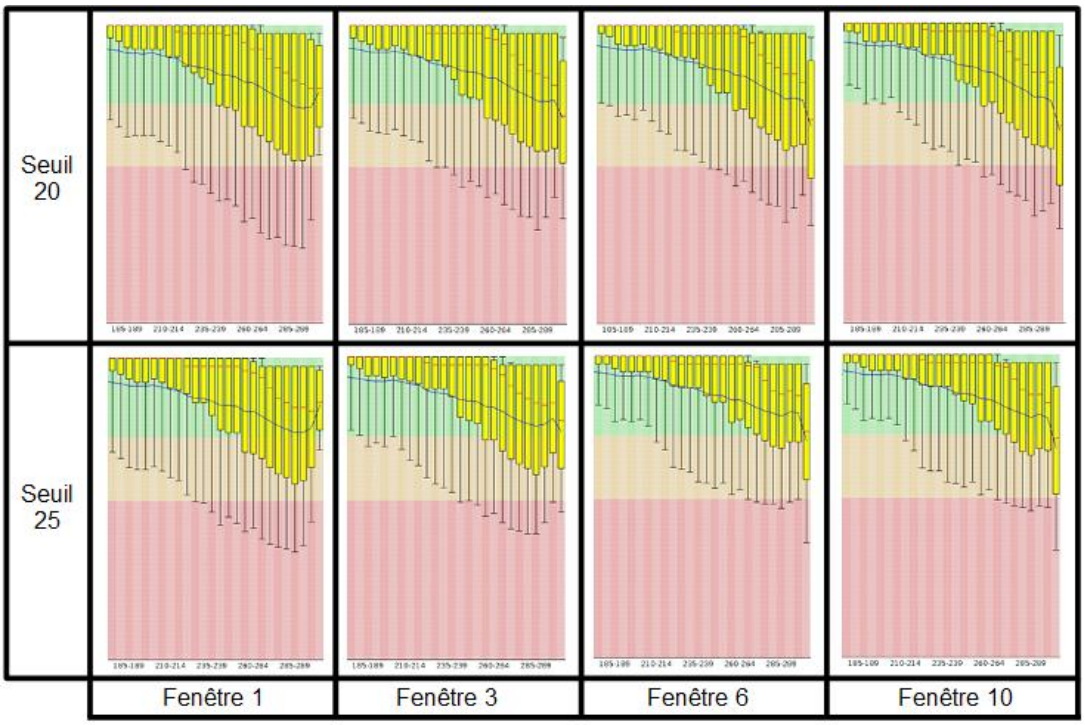


Figure 30 : Schéma comparatif entre différents paramètres de fenêtre glissante. *La taille est comprise entre 1 et 10 et le seuil de qualité de base entre 20 et 25. Exemple sur le read 2 de l'accession 'Georges Delbard'. (2015). N. Jacquier*

On observe en comparant les fenêtres 2 à 2 que le seuil de qualité de 25 réduit de façon plus importante les écarts entre les quartiles et entre les déciles et augmente les valeurs des moyennes et des médianes. De plus, on constate que la fenêtre 1 a un impact plus limité sur le nombre de read trimmé. En effet, seules les bases des dernières positions sont éliminées. En revanche, la différence entre les fenêtres 6 et 10 n'est pas notable hormis un trimming légèrement moins élevé au niveau des dernières bases car la valeur du premier décile et du premier quartile de la fenêtre de taille 6 est légèrement plus élevée que pour la fenêtre 10. Cela est vraisemblablement dû au fait que la fenêtre glissante de taille 10 s'est stoppée plus tôt que la 6 et donc il reste quelques bases de moins bonnes qualité en fin de séquence du read.

Pour confirmer plus en détail ces observations, le graphique ci-contre (Fig.31) montre l'évolution des premiers quartiles en fonction des différentes fenêtres glissantes (les données brutes des premiers déciles ne sont pas disponibles dans le fichier de sortie). Ces résultats indiquent que 75% des bases pour une position donnée ont une valeur de qualité plus élevée pour la fenêtre 6 que la 10 et la 3.

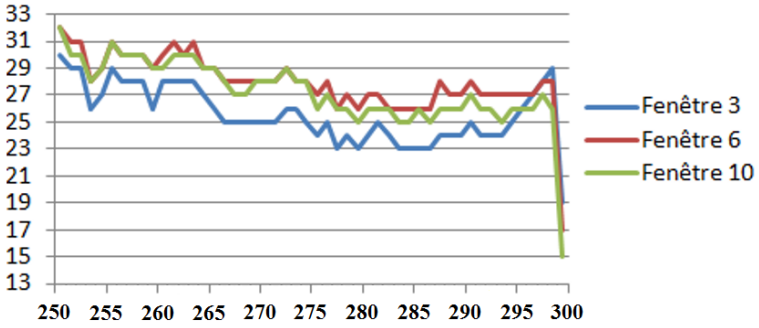


Figure 31: Comparaison de l'évolution de la valeur des premiers quartiles en fonction de la position des bases au sein du read pour 3 tailles de fenêtres glissantes différentes. *(2015). N. Jacquier*

A la valeur de qualité des bases s'ajoute la profondeur de chacune d'entre-elles, c'est-à-dire le nombre de bases qui sont présentes en une position donnée. Ainsi lorsque les reads ne sont pas trimmés, la profondeur est constante. En revanche, plus le trimming est important, moins il y a de bases présentes en fin de read. On perd alors des informations, même si elles sont de mauvaise qualité. Il faut donc faire un compromis entre l'élimination des bases et leur représentation.

On constate (Fig.32) qu'en fin de read, on a une réduction de la profondeur d'environ 50% par rapport aux données non trimmées. De plus, quel que soit la taille de la fenêtre, la profondeur ne diminue pas de plus de 10% pour les 250 premières bases. Les résultats entre les fenêtres 6 et 10 sont proches et ce sont principalement les 20 dernières bases pour lesquelles des données seront manquantes.

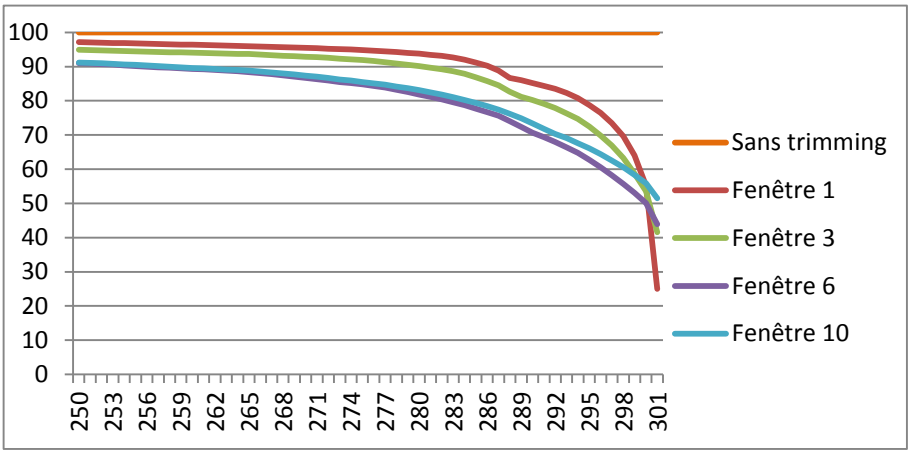


Figure 32: Comparaison de la profondeur des données en fonction des différentes positions des bases pour 4 tailles de fenêtres différentes et une valeur seuil de 25. *Le seuil 20 n'est pas présenté pour des raisons de clarté et seule la profondeur des 50 dernières bases est représentée. (2015). N. Jacquier*

Les paramètres de qualité les mieux adaptés pour notre étude sont donc une fenêtre de 6 et une valeur seuil de 25. Cependant, à l'issue de cette première étape de trimming, il demeure en fin de reads des bases de mauvaise qualité. Nous avons donc réalisé un second nettoyage des reads avec une fenêtre de taille 1 et une valeur seuil de 25, c'est-à-dire que les dernières bases du read ont été éliminées si elles n'étaient pas au-delà de cette valeur. (Annexe IX).

3.3. Le Mapping

3.3.1. Objectif général

Comme nous l'avons évoqué auparavant, le mapping consiste en l'alignement le long de la séquence de référence des reads forward et reverse, que nous qualifierons de couple ou reads pairés. Pour cette étape, nous avons dû choisir le logiciel le mieux adapté à nos besoins qui sont :

- Gestion d'une longueur de read de 300 bases
- Détection à la fois des insertions et des délétions (Indels) par rapport à la séquence de référence
- L'association des deux reads appartenant bien au même couple
- Le respect de la « taille d'insertion de fragment attendu » ou « insert size » (Fig.33). Il s'agit pour des données Illumina pairées de la différence entre les extrémités 5' des deux reads soit la taille des amplicons attendus, amorces comprises [39].

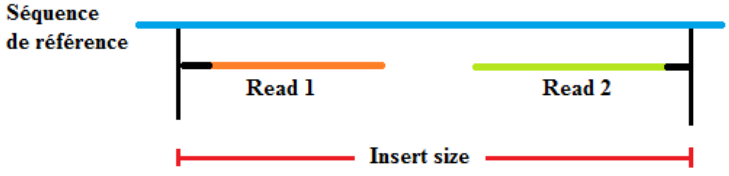


Figure 33 : Représentation de l'insert size. *(2015). N. Jacquier*

Pour ce faire, nous avons comparé différents logiciels proposés sur l'instance Galaxy d'après des données bibliographiques, puis nous les avons testés à l'aide de nos données brutes pour pouvoir analyser les fichiers de sortie.

3.3.2. Comparaison des logiciels de mapping

Trois outils sur l'instance Galaxy sont proposés (du plus ancien au plus récent) : BWA [40], Bowtie et Bowtie2 [34].

Des analyses de performance [41] ont montré que Bowtie2 est le logiciel le plus rapide pour le traitement des données et qu'il utilise le moins de mémoire. Il est suivi de Bowtie [34] puis BWA. Par rapport à Bowtie, ce sont surtout pour des reads d'une longueur supérieure à 50 bases que Bowtie2 est le plus rapide [41]. Ceci n'est pas à négliger vu le grand nombre de données que nous avons à traiter. Concernant la réalisation du mapping en lui-même, les algorithmes utilisés sont différents entre BWA et les Bowtie [42]. De plus, Bowtie [43] prend en compte la qualité de chaque base lors de l'alignement et est plus tolérant avec les bases de mauvaise qualité. En revanche, BWA fonctionne en éliminant les bases de mauvaise qualité et affiche une qualité de mapping globale (Medina et al., 2012).

A l'inverse de Bowtie, Bowtie2 et BWA ont l'avantage de gérer les alignements discontinus, c'est-à-dire les délétions et insertions. De plus, Bowtie 2 n'a pas de nombre de délétions limité et permet un alignement des bases face à des bases ambigües de la séquence de référence, notamment les N, qui symbolisent des bases non connues. [34] Ce dernier point est à ne pas négliger car les séquences de Old Blush contiennent parfois des N ou des lettres indiquant plusieurs bases possibles sans qu'un compromis ait pu être trouvé [44], (Annexe 12). Enfin, nous avons pu constater que près de 30 paramètres de mapping sont disponibles sur Bowtie. Cela peut permettre un mapping optimisé mais en cas de mauvais réglages entraîner un mapping erroné.

Au vu de tous ces critères, Bowtie 2 était le logiciel le plus adapté de par sa rapidité de mapping pour des fragments de 300 bases mais surtout parce qu'il gère les indels et la présence de bases ambigües.

3.3.3 Choix des paramètres

Certains paramètres ont été choisis et fixés :

- Le mode d'alignement est le mode « end-to-end ». Celui-ci est plus approprié lorsque les reads ont été élagués des bases de mauvaise qualité et qu'ils ne contiennent pas les adaptateurs. En effet, ce mode permet d'aligner l'ensemble de la séquence contrairement au mode « local » qui permet d'éliminer des bases aux extrémités pour améliorer l'alignement qui peut n'être que partiel et qui est donc plus adapté à des reads non trimmés [45].
- L' « insert size » maximale a été fixée pour chaque gène candidat en prenant la valeur correspondante à la taille du plus grand amplicon produit possible pour ce gène.
- Les indels ne sont pas autorisés dans les 20 premières bases du read, ces dernières correspondant aux amorces (Annexe IX).

3.3.4 Vérification des résultats du mapping

Pour déterminer si les résultats obtenus à l'issue du mapping étaient corrects, nous avons réalisé le mapping sur deux jeux de données ('Old Blush' et 'Georges Delbard') afin de repérer

les mauvais paramétrages [46]. Les données disponibles dans le fichier de sortie (Fig.34) pour chaque individu sont [47] :

1- QNAME : Nom identique pour un même couple de reads (forward et reverse).

2 - FLAG : Code qui indique diverses informations de mapping (cf plus loin).

3 - RNAME : Nom de la séquence de référence sur laquelle le read a mappé.

4 - POS : Position sur la séquence de référence de la 1ère base du read ayant mappé avec la séquence de référence.

5 - MAPQ : Qualité du mapping.

6 - CIGAR : Code indiquant les régions du read mappées, les insertions et les délétions.

7 - MRNM : Nom de la séquence de référence sur laquelle le 2nd read du couple a mappé.

8 - MPOS : Position de la 1ère base du 2nd read du couple ayant mappé avec la séquence de référence.

9 - ISIZE : Taille de la séquence, soit nombre de bases entre la première et la dernière base des reads pairés. Elle correspond à la taille des amplicons, amorces comprises.

10 - SEQ : Séquence du read ayant mappé.

11 - QUAL : Indication du score de qualité de chaque base du read ayant mappé.

12 - OPTIONS : Résultat des différentes options choisies.

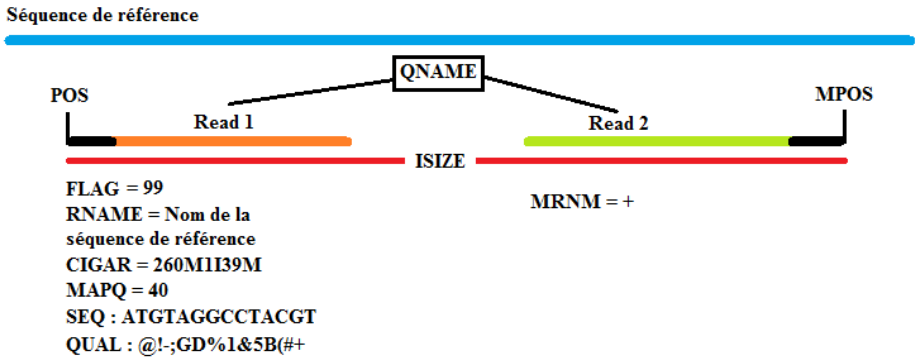


Figure 34 : Informations disponibles à l'issue du mapping pour chaque read. *Exemple des informations de sortie prises par rapport au read 1. Le QNAME est la seule information commune et la seule donnée spécifique au read 2 est le MRNM. (2015). N. Jacquier*

Les flags sont particulièrement intéressants car ils permettent de vérifier les résultats de mapping et donc les paramètres que nous avons choisis.

Un numéro de flag est le résultat de la somme de différentes valeurs qui correspondent chacune à une information particulière [47]. Celles-ci sont issues d'une succession de codages binaires : si la condition testée est respectée, la séquence binaire reçoit un 1, si elle est fausse un 0. Le 1 est ensuite converti en un nombre associé à l'information codée. Par exemple, la séquence binaire 1100011 signifie 64+32+0+0+0+2+1=99 et traduit donc le flag 99. Chaque read a ainsi un code flag unique, soit 2 flags différents par couple : 1 pour le read forward et 1 pour le reverse.

Les informations testées et associées aux valeurs permettant de calculer les Flags de reads pairés de Bowtie sont les suivants [48]+[49] :

- **1** = Le read fait partie d'un couple
- **2** = Les deux reads du couple sont mappés
- **4** = Le read n'est pas mappé
- **8** = Le 2nd read du couple n'est pas mappé
- **16** = Le read a mappé sur le brin reverse de la séquence de référence
- **32** = Le 2nd read du couple a mappé sur le brin reverse de la séquence de référence
- **64** = Le read mappé est le forward
- **128** = le read mappé est le reverse

Les flags observés durant l'étude sont résumés ci-dessous (Fig.35), (Annexe XIII, tableau complet):

FLAG	flags	pair	itself		mate	
One of the mate is unmapped						
73	1+8+64 73	1	map	+	unmap	
133	1+4+128 133	2	unmap		map	+
69	1+4+64 69	1	unmap	+	map	+
137	1+8+128 137	2	map	+	unmap	+
Both unmapped						
77	1+4+8+64 77	1	unmap	+	unmap	+
141	1+4+8+128 141	2	unmap	+	unmap	+
mapped in correct orientation and within insert size						
99	1+2+32+64 99	1	map	+	map	-
147	0+1+2+16+128 147	2	map	-	map	+
83	1+2+16+64 83	1	map	-	map	+
163	1+2+32+128 163	2	map	+	map	-
mapped uniquely but wrong insert size, and could reside in different contigs						
81	1+16+64 81	1	map	-	map	+
161	1+32+128 161	2	map	+	map	-
97	1+32+64 97	1	map	+	map	-
145	1+16+128 145	2	map	-	map	+

Figure 35 : Les principaux flags rencontrés. *+ signifie que le read a mappé sur le brin forward de la séquence de référence et le - sur le brin reverse. 99, 147, 83 et 163 sont les flags que l'on souhaitent obtenir car ils correspondent au fait que les 2 reads du couple ont mappés sur la séquence de référence en respectant l'insert size maximale spécifiée. [48] (2015). N. Jacquier*

Analyse des données de sortie :

Dans notre étude, les reads faisant partie d'un couple et l'insert size ayant été spécifiée, les flags attendus et correspondant à un bon mapping sont 99 et 147 si les reads ont mappés sur le brin forward de la séquence de référence, 83 et 163 sinon [50].

Une observation des flags de sortie a permis de déceler une erreur de paramétrage en amont du mapping. En effet, des flags, 73, 133, 141 … (visibles Fig.35) étaient visibles, indiquant alors que soit les reads n'avaient pas mappés, soit ils avaient mappés avec une mauvaise « insert size ». Pour comprendre l'origine de ces résultats, les QNAME de chacun des reads pairés ont été récupérés et leur flag comparés. Ainsi, un read forward pouvait avoir un code de 73 et un read reverse un code de 141. Or 73 indique que seul le read forward a mappé alors que le flag 141 indique que les deux reads du couple ne sont pas mappés. Ce sont des résultats contradictoires pour deux reads d'un même couple. Ils indiquent donc un mauvais mapping qui génère des données erronées.

Pour déterminer la source du problème, il a fallu se demander comment fonctionnait le logiciel pour associer les deux reads d'un couple l'un avec l'autre. En effet, le QNAME semblait être le seul paramètre permettant au logiciel d'associer les bons reads ensemble. Cependant, en observant les données de séquençage, il s'avère que les read forward et les reads reverse sont dans deux fichiers séparés mais ils sont classés dans le même ordre au sein de ces fichiers. Il est donc possible que les reads soient simplement associés ligne par ligne. En réexaminant tous les paramètres de l'ensemble de nos logiciels, il s'avère que la case « conserver les reads d'une longueur de 0 » n'avait pas été cochée et de ce fait, certains reads étaient éliminés dans un fichier ou dans l'autre mais sur des lignes différentes. Cela créait donc un décalage dans l'association des reads deux à deux. Après rectification, les flags en sortie étaient bien 99, 147, 83 et 163 hormis une dizaine (flags 73 ou 133 par exemple) qui sont issus d'une amplification non souhaités puisque seul l'un des deux reads est mappé.

La qualité du mapping :

A l'issue du mapping, les différents scores de qualité observés sont : 3, 8, 23, 24, 40 et 42. Ces valeurs indiquent un alignement unique du read [51]. Plus la valeur est élevée, moins il y a de zones d'insertion et délétion. Si le MAPQ est de 0, le nombre de zones avec indels est supérieur à 5 et la séquence n'est pas rapportée [52].

Enfin, une vérification des données de l'alignement en lui-même (Fig.36), traduite par le code Cigar, a été réalisée. Les codes Cigar sont composés de 3 types d'informations :

M : Match/Mismatch. Indique que la base a mappé sur la séquence de référence qu'elle soit identique ou différente de celle-ci.

I : Indique une insertion au sein du read par rapport à la séquence de référence.

D : Indique une délétion au sein du read par rapport à la séquence de référence.

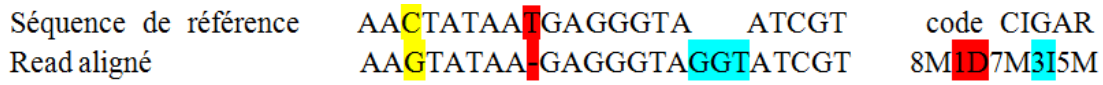


Figure 36 : Représentation des indels indiqués par le code Cigar entre la séquence du read et la séquence de référence. *Les insertions sont en bleu, les délétions en rouge et les mismatch en jaune. (2015) N. Jacquier.*

Des données de read alignées contre la séquence du gène *NUDX* ont été récupérées puis alignées sur Geneious. En effet, des informations de polymorphisme étaient disponibles pour cette séquence. Cela a permis de constater que les indels étaient bien placées ainsi que les SNP.

3.4. Le contigage intra-séquence

A l'issue du mapping, un script, mis au point pour des analyses de pommier, devait être utilisé afin de réaliser un contigage intra-séquence, c'est-à-dire la jonction entre les deux reads d'un même couple en un seul que l'on appellera read contigué. De même, ce script devait permettre d'effectuer un dosage d'haplotypes, soit le comptage du nombre d'allèles différents observés et de l'occurrence de chacun des haplotypes. Ceci a pour but de déterminer les effectifs alléliques de chaque individu. Cependant, il s'est avéré que les fichiers .sam nécessaires au bon fonctionnement du script n'étaient pas les mêmes entre ceux générés par Galaxy ou par le logiciel CLC, qui dispose d'un programme de mapping qui lui est propre. C'est en effet ce dernier logiciel, payant, qui avait fourni les données d'entrée du script et pour lesquelles celui-ci avait été conçu. Ainsi, certaines informations manquantes dans nos fichiers ont bloqué l'exécution du script.

Un nouveau programme a alors été mis au point pour réaliser ces opérations. Pour cela Excel 2007 a été utilisé car il permet à la fois l'utilisation de fonctions prédéfinies et aussi la réalisation de macros, c'est-à-dire de programmation.

L'avantage de la conception d'un code est qu'il s'adapte à nos besoins. Ainsi, certains points vont pouvoir être pris en compte contrairement au script initial:

- Les amorces spécifiques vont pouvoir être retirées ce qui n'était pas le cas jusqu'à présent.
- Les erreurs de séquençage vont pouvoir être corrigées.
- La ploïdie va pouvoir être prise en compte lors du dosage d'haplotypes de chaque individu (étape non achevée actuellement).

3.4.1. Le fichier Excel

Le fichier mis au point permet le traitement quasi instantané de 500 reads à la fois. Une automatisation permettra de traiter à la chaîne tous les reads pour une séquence de référence

donnée. Ce programme est donc spécifique pour chacun des gènes candidats mais le principe général reste le même. En effet, il faudra simplement adapter certaines conditions telles que la suppression des amorces spécifiques. Les données d'entrée utilisées sont celles issues du mapping de Bowtie2 et filtrées sur les flags 83, 99, 147 et 163. Cela élimine ainsi tous les reads non mappés ou qui auraient amplifié d'autres portions du génome.

3.4.2. Les différentes stratégies réalisées par le programme de contigage

Gestion des données d'entrée

Dans un premier temps, les données en entrée (12 colonnes avec les paramètres cités en partie 3.3.4) sont triées par QNAME en ordre alphabétiques afin que les 2 reads d'un même couple soient sur deux lignes adjacentes du fichier. Grâce aux flags, le read forward est toujours placé avant le reverse. Ainsi, le logiciel travaillera sur les lignes pour associer les reads entre eux et non à partir de leur identifiant, ce qui simplifiera le traitement des données. Ensuite, seules les informations de la séquence de référence, le POS, le code Cigar, la séquence du read et la séquence des valeurs de qualité sont conservés. Puis, l'ensemble des bases des séquences des reads sont séparées de sorte à n'en avoir qu'une seule par colonne.

Le code cigar et la stratégie de traitement des séquences des reads

Ensuite, le code Cigar est décomposé pour obtenir dans des colonnes différentes le nombre et la lettre. On obtient ainsi par exemple : 130 | M | 3 | I | 51 | M | 2 | D | 64 | M pour le code 130M3I51M2D64M.

Par la suite, il va falloir aligner nos reads sur la séquence de référence pour pouvoir ensuite les contiguer. Cette étape est essentielle car les bases des zones de chevauchement doivent être placées dans la colonne correspondant à leur position par rapport à la séquence de référence et ce pour ne pas commettre d'erreur lors du contigage. Or les reads présentent des indels par rapport à la séquence de référence, ce qui peut entraîner des décalages par rapport à celle-ci. Il faut donc aligner la séquence du read sur la séquence de référence en comblant les délétions et en enlevant les insertions.

Il a donc fallu calculer les différentes positions des insertions et des délétions au sein des séquences des reads (Fig.37). En effet, lorsqu'une insertion est indiquée, cela signifie que le read mappé a un certain nombre de bases en plus par rapport à la séquence de référence et qu'il faut donc retirer. Pour ne pas perdre cette information, les bases extraites sont conservées sur une autre feuille de calcul et la première base suivant l'insertion est enregistrée puis modifiée par une lettre différente de celles des bases c'est-à-dire « d », « f », « h » et « j » si les insertions sont sur le read forward et « e », « p », « i » et « k » si les insertions sont sur le read reverse (Fig.37, « d » en position 20). Les différentes lettres indiquent s'il s'agit de la première insertion (« d » ou « e »), la seconde (« f « et « p »)… Cette stratégie nous permet de conserver l'information de la séquence insérée et de la base qui suit, tout en ne compromettant pas l'alignement du read contre la séquence de référence. Une valeur de la qualité de 41 est alors attribuée à une insertion pour éviter la perte de ses informations. A l'inverse, les zones de délétions doivent être repérées et un X est ajouté pour chaque base supprimée afin ne pas décaler l'alignement.

Position des bases	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Séquence originale du read	A	T	T	G	C	G	T	A	A	A	G	C	T	T	G	A	C	A	C	T	G	T
Code cigar	10M										3D			6M			3I			3M		
Position extraction	1 à 10										0			de 11 à 16			de 17 à 19			de 20 à 22		
Nombre de X	0										3			0			0			0		
Insertion	0										0			0			3			0		
Séquence finale du read	A	T	T	G	C	G	T	A	A	A	X	X	X	G	C	T	T	G	A	**d**	G	T
Séquence de référence	A	T	C	G	C	G	T	A	A	A	G	T	C	G	C	T	T	G	A	T	G	T

Figure 37 : Schéma illustrant les calculs à réaliser pour extraire les zones d'insertions et ajouter des X pour les délétions. *(2015). N. Jacquier*

En ce qui concerne les positions de bases sur la séquence, il est important de prendre en compte les indels afin d'aligner le read sur la séquence de référence sans erreurs ni décalages, d'ajouter les délétions au bon endroit et d'extraire les bonnes insertions. En effet, lorsque la première zone d'insertion est détectée, sa position de départ est connue grâce au nombre de bases indiqué par le code Cigar avant le M. Ainsi, si le code Cigar est 53M3I, la « zone matchée »[1], se situe entre la première base du read et la 53ème tandis que l'insertion est en position 54. En revanche, pour déterminer la position des zones d'insertions ou délétions suivantes, des calculs doivent être opérés (Fig.37, ligne 4). Il s'agit du calcul de la somme des nombre de bases correspondant aux M et aux I. En effet, si l'on a une insertion, le nombre de bases de celle-ci doit être pris en compte pour pouvoir récupérer les bases de la prochaine zone ayant matchée. Si c'est une délétion, le nombre du code Cigar permet de connaître le nombre de X à ajouter, car il identifie des bases manquantes par rapport à la séquence de référence. Néanmoins, le nombre de délétions ne doit pas être pris en compte pour la position des prochaines bases matchées, puisqu'elles ne sont pas présentes au sein du read d'origine mais introduite seulement après. Compter les délétions nous ferait donc extraire les mauvaises bases pour la suite.

A la fin de ces opérations, des N sont ajoutés avant et après le read de sorte qu'ils aient tous la même longueur quel que soit leur position et donc qu'ils s'alignent parfaitement sur la séquence de référence. Le nombre de N s'ajoutant avant le read est calculé grâce à l'opération POS-1. Ainsi, si POS=10, alors on ajoute 10-1=9 N. Le nombre de N à ajouter à la fin correspond lui au nombre de bases de la séquence de référence auquel on enlève la longueur du read final et le nombre de N ajoutés avant. Ainsi, on a un aperçu de tous les reads positionnés le long de la séquence de référence, quelques soient les couples d'amorces (Fig.38).

Référence	A	C	G	A	C	G	C	C	A	A	T	T	T	C	T	C	C	A	G	C	C
Position	**272**	**273**	**274**	**275**	**276**	**277**	**278**	**279**	**280**	**281**	**282**	**283**	**284**	**285**	**286**	**287**	**288**	**289**	**290**	**291**	**292**
Read	**N**	**N**	G	A	C	G	**a**	C	A	A	T	T	T	C	**X**	C	C	A	G	**N**	**N**

Figure 38 : Séquence finale d'un read issue du traitement des indels et alignée sur la séquence de référence. *La position 274 correspond à la POS. (2015). N. Jacquier*

Gestion des valeurs de qualité

En parallèle du travail sur les séquences des reads, les mêmes opérations sont réalisées avec les séquences de qualité mais à cela s'ajoute une phase de conversion. En effet, chaque valeur de qualité est associée à un symbole, or il est nécessaire de les convertir en nombre pour pouvoir réaliser des comparaisons de qualité. Les données de conversion reprennent celles présentées en partie 3.2.2. Cependant, certains symboles n'étant pas reconnus par les fonctions Excel, ils ont été changés avant la conversion (Annexe XIV).

[1] Traduit le fait qu'une zone déterminée de la séquence du read est totalement alignée face à la séquence de référence. Cette « zone matchée » peut contenir des SNP mais pas d'indels. Elle est donc située entre deux indels ou entre une amorce et un indel si elle est en début de read.

3.4. Le contigage intra-séquence

On obtient donc à l'issu de ces traitements de données deux feuilles avec sur l'une les reads et sur l'autre les valeurs de qualité des bases correspondantes.

La stratégie de contigage intra-séquence

On peut alors réaliser un contigage deux lignes par deux lignes en mettant en parallèle les bases et leur qualité. Ainsi, une séquence consensus est créée en conservant : soit la base commune aux deux reads (au niveau d'une zone de chevauchement), soit en choisissant la base du read dont la valeur de qualité est la plus élevée. Les N ne sont donc pris en compte que lorsque la taille de l'amplicon est supérieure à 600 ou lorsque les reads avaient été raccourcis par le quality trimming. On obtient donc des reads contigués, pouvant contenir des N en leur centre, et toujours alignés sur la séquence de référence du type :
NNNNNNATTAGCCTCAA**d**ATGCTTAGT**XX**ACAAGATC**i**GTATTAGGGTC**NNNNNN**

Le retrait des amorces

A l'issue du contigage, les amorces des contigs, initialement aux extrémités de chaque read, sont retirées. En effet, malgré un mismatch entre l'amorce et la matrice ADN, il peut parfois y avoir amplification. Le read résultant de cette amplification présente la séquence de l'amorce et pas celle de la matrice. Dans la perspective des analyses de diversité, ce biais technique est éliminé en retirant les amorces des reads contigués. Pour cela, les emplacements des amorces sur la séquence de référence sont identifiés et une fonction permet d'analyser si ces mêmes séquences sont présentes au sein des reads contigués. Si c'est le cas, elles sont remplacées par des N et le nom du couple d'amorces correspondant est associé au read.

La correction des erreurs de séquençage

Enfin, nous avons mis en place une stratégie pour nettoyer les erreurs de séquençage des reads. Pour ce faire, pour chaque position de la séquence de référence, les proportions des bases (A, T, G et C) et des indels sont calculées en ne prenant pas en compte les bases non connues N. Lorsque la proportion d'une base en une position donnée ne dépasse pas une fréquence seuil, elle est remplacée par la base la plus fréquente. Les reads avec des erreurs sont donc corrigés et non pas supprimés. Ainsi, conserver l'ensemble des reads augmente la profondeur pour une base donnée et diminue le risque de mal estimer la présence d'une base par rapport à une autre.

Le seuil de 5% a été choisi en raison de la ploïdie de nos rosiers. En effet, au maximum, les rosiers de nos échantillons sont hexaploïdes. Dans ce cas, il peut y avoir jusqu'à 6 allèles par gène donc au minimum, ils sont représentés par 16% des reads. Cependant, une amplification compétitrice entre allèles peut diminuer cette valeur. Le seuil devra donc être abaissé. Le taux d'erreur du séquençage MiSeq est estimé à 0.8% [53] mais pour des reads d'une longueur supérieure à 200pb, ce taux est de 1% au-delà de 100pb et entre 4 et 5% en fin de read (Quial et al., 2012). De plus les erreurs sont plus importantes lorsque le motif GGC est rencontré et plus particulièrement la succession GGCGGG. En observant les résultats des proportions de chaque base pour un individu tétraploïde pour le gène *F3'H*, il a été observé que les bases pouvant résulter d'erreur de séquençage, ou issue d'une éventuelle erreur de PCR ne dépassaient pas les 3%. Nous avons donc placé le seuil à 5%, pour ne pas éliminer trop d'informations mais ne pas considérer une base très faiblement représentée comme un SNP, notamment en fin de read.

Ce nettoyage des données facilitera la prochaine étape du traitement qui est en cours de réalisation : le dosage d'haplotypes.

Partie 4 : Discussion

4.1. Bilan sur la qualité des amorces

Comme nous avons pu le voir (partie 3.1), le design de nos amorces a été optimisé en fonction de différents critères dont certains pouvaient représenter une source d'échec de l'amplification. Le principal risque est dû à une grande incertitude quant à la présence de SNP au sein des amorces et donc un risque de faible ou non amplification du gène candidat ou bien d'une amplification d'un autre segment d'ADN, ou encore d'une amplification compétitive entre allèles.

L'EPGV d'Evry nous a communiqué une liste de 10 couples d'amorces (Tab.5) pour lesquelles après un alignement réalisé sur CLC aucun read n'avait mappé pour 'Old Blush', qui est pourtant le rosier duquel sont issues nos séquences de référence, utilisées pour la création des amorces. Or, une analyse de notre jeu de données issu du traitement sur Galaxy nous indique la présence de reads, même si certains sont présents en faible quantité.

Tableau 5 : Tableau comparatif des informations connues pour les couples d'amorces n'ayant pas fonctionné pour 'Old Blush' selon l'EPGV et nos données. *La profondeur communiquée par l'EPGV était qu'aucun read n'avait mappé pour ces couples sur 'Old Blush'. En rouge, les Tm inférieures à 59°C et donc n'appartenant pas à la fourchette de Tm idéales. (2015). N. Jacquier*

Amplicons			Amorce forward		Amorce reverse	
Nom	**Taille**	**Profondeur maximale observée pour 'Old Blush' après mapping par Bowtie2**	**Tm**	**Remarque**	**Tm**	**Remarque**
AG_E1E1	221	671	**58,57**		59,71	
BRC1_E1E1b	521	1030	59,9		59,8	
F3'H_E1E2	760	184	59,8		59,7	
FT_E1E4	741	513	**58,88**	Max Poly-X=5	**58,07**	
FT_E2E3	421	325	59,6		**58,79**	
PAAS_E1b	521	516	60,05	Max Poly-X=4	**57,24**	
ZDS_3UE13	596	686	**58,74**		59,2	
ZDS_E1E2	707	531	59,8		59,4	
ZDS_E3E6	933	169	59,48	Max Poly-X=4	59,4	
ZDS_E7E8	395	782	60,4		59,69	1 SNP

Des explications possibles à ces différences de profondeur de reads entre celles obtenues par Evry (CLC) et nos données (Bowtie2) pourraient être une sélection plus stricte des reads sur CLC [54] en raison du fait qu'au-delà d'un certain nombre d'indels, il est possible que le read ne soit pas reporté et donc sur des segments d'ADN ayant beaucoup de variabilité, les reads mappés ne sont pas présents dans les fichiers de sortie. De plus, CLC réalise un « local alignement » [91] et aligne donc seulement les zones qui matchent correctement, en laissant le reste du read non aligné. De ce fait, les reads mappés sont indiqués comme ayant un mauvais alignement. Le mode d'alignement n'est pas le plus adapté à nos données.

En outre, les différences de profondeur peuvent s'expliquer par une hybridation plus difficile de l'amorce en raison de la présence de SNP non identifiés au sein de sa séquence. Cela pourrait être le cas des couples ZDS_E3E6 et F3'H_E1E2 qui ont une très faible profondeur. De même, un nombre important de bases identiques adjacentes dans l'amorce (Max Poly-X) et une Tm inférieure à 59°C pourraient diminuer l'amplification. Enfin, la taille des amplicons n'est pas à négliger car s'ils sont de taille inférieure à 300 bases, il y a un risque qu'ils soient éliminés durant l'étape de purification. Cette donnée ayant été communiquée tardivement, le couple AG_E1E1 avait été créé avec une longueur de 221 bases. La profondeur reste très convenable

mais lors de l'amplification des autres rosiers de l'échantillon, il est possible qu'un SNP soit présent dans l'amorce et la profondeur risque de chuter de façon importante.

Il aurait donc été plus pertinent de vérifier l'amplification de chaque amorce avant de lancer les runs. En effet, une amplification individuelle sur les 8 individus tests était initialement prévue. Cependant, en raison de contraintes de temps et pour vérifier le bon fonctionnement des machines de séquençage, les amorces sont passées directement dans le run test de PCR access Array puis dans le séquenceur. Or, pour l'étape de séquençage, 3 autres runs avec l'ADN de 48*3 de nos individus sont passés en même temps car le prix de l'opération pour un seul run est bien trop élevé.

Ceci avait un avantage car on a pu constater que plusieurs individus ont pu être amplifiés en faible quantité mais que celle-ci était suffisante pour le séquençage.

L'inconvénient de cette démarche est que des couples d'amorces fonctionnant mal et qui auraient pu être détectés avec un test préalable ont déjà été utilisés pour le séquençage de 3 jeux de données.

4.2. La stratégie de design des amorces vis-à-vis du contigage

A l'issue du quality trimming (partie 3.2), certains reads sont fortement raccourcis en raison de bases de mauvaise qualité à leur extrémité 3'(séquences bleues sur le schéma, Fig.39 B). Or, cela réduit l'optimisation du contigage intra-séquence et la connaissance de l'ensemble de la séquence du fragment. En effet, à la fin de cette étape, on se retrouve avec des reads ayant des N au sein des séquences car la zone de chevauchement de 50pb initialement prévue n'a pas été conservée à l'issue du trimming (fig.39 C). Or, nous avons vu que les 100 dernières bases sont touchées lors de cette étape. Ainsi, sur les 600 bases que l'on peut obtenir, il en faudrait au moins 100 pour la zone de chevauchement. Pour le design des prochaines amorces, il serait conseillé de réduire, lorsque cela est possible, la longueur de nos amplicons en conservant une zone de chevauchement de 100 bases pour minimiser les pertes d'informations.

De la même façon, la zone de chevauchement inter-séquence devrait être augmentée, pour faciliter le contigage entre amplicons adjacents, d'autant plus que les régions des contigs correspondant aux amorces ne peuvent être utilisées pour le contigage car elles sont non informatives. Cependant, ceci sera plus difficilement réalisable car certains exons sont très petits et la zone de chevauchement ne pourra donc pas être étendue. Une solution possible serait de créer des amorces qui génèrent des amplicons de grande taille (Fig.39 A') mais dont les 200 premières bases de chacun des reads se situent au niveau de deux couples d'amorces différents. En combinant les informations de ces trois couples, le contigage inter-séquences serait alors possible et on pourrait également récupérer des données pour les bases manquantes (Fig.39 C et C').

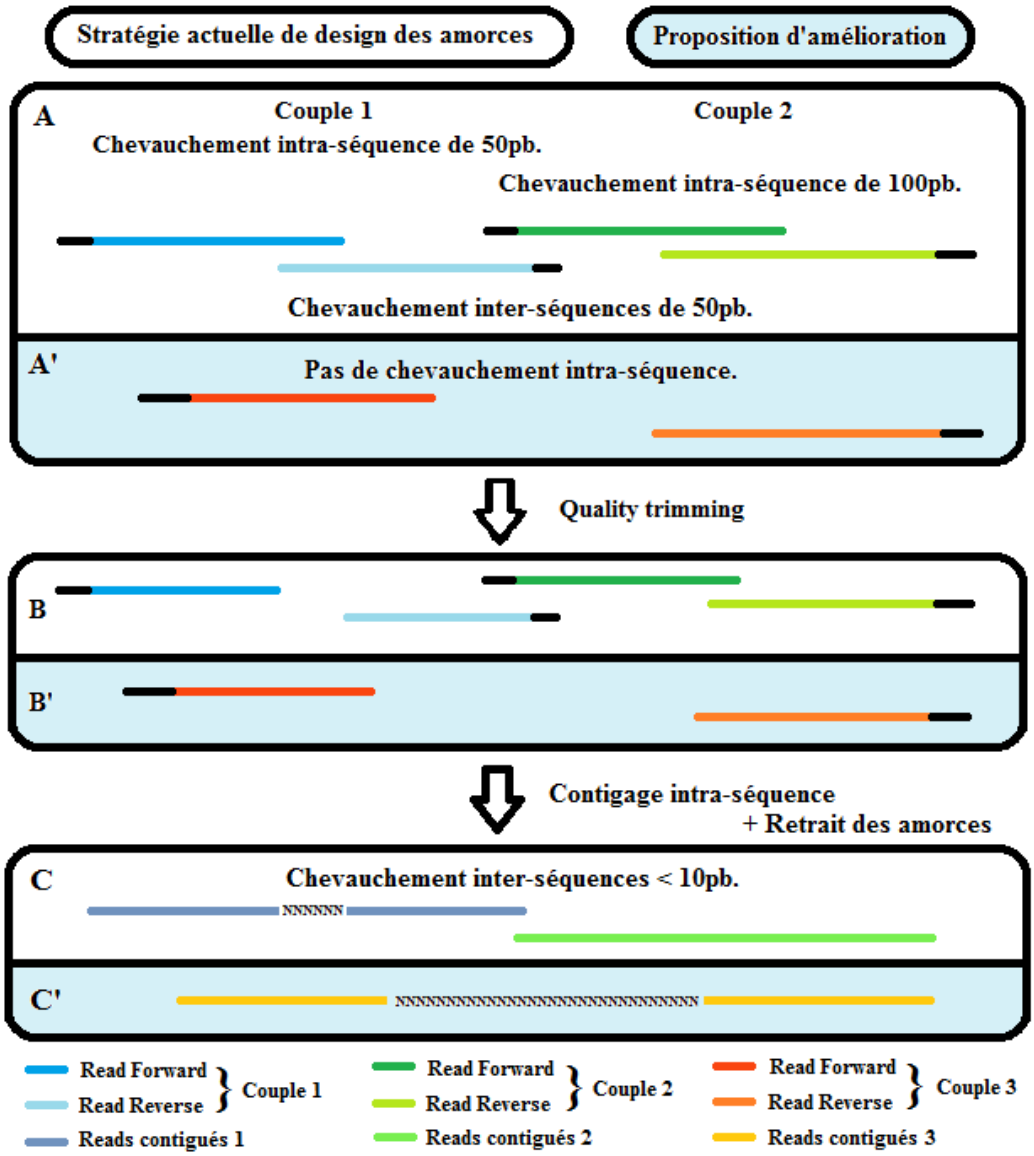


Figure 39 : Schéma bilan de la stratégie actuelle du design des amorces et proposition d'amélioration. *L'objectif de cette proposition est de récupérer des données manquantes (bases N non identifiées) et faciliter le contigage inter-séquences. Pour ce faire, de nouveaux couples d'amorces sont mis au point (orange) en les plaçant à cheval sur deux zones déjà amplifiées. Cela permettra de combiner les hapltotypes dosés pour chaque couple et en déduire lesquels sont à associer ensemble. (2015) N. Jacquier.*

4.3. Automatisation de Galaxy pour les analyses en série

Un problème majeur de l'instance Galaxy est le manque d'automatisation de l'enregistrement des données de sortie. En effet, il est généré un fichier par gène et par individu soit dans le cas de l'étude complète plus de 30 000 fichiers. Il va donc falloir minimiser le nombre d'enregistrements à réaliser sur l'instance Galaxy. Une première solution consiste en la concaténation des résultats des 48 couples d'amorces dans un même fichier par individu. Ainsi, il n'y aura « plus que » 384 enregistrements à réaliser pour les 2 lots de 48 amorces. De plus, une autre lacune de Galaxy est que l'on ne peut pas renommer automatiquement le fichier de sortie. Le passage par un manipulateur est donc toujours nécessaire. La solution proposée du concaténage est à l'heure actuelle la seule qui permette un compromis entre la modification du nom du fichier et la minimisation des enregistrements.

4.4. Analyses post-mapping sous Excel

Concernant le fichier Excel mis en place, c'est essentiellement le fait d'avoir déjà des connaissances sur les fonctions qui pouvaient être utilisées et quelques notions de code VBA qui a fait que ce logiciel a été choisi. De plus, la découverte de l'incompatibilité des données avec le script ayant eu lieu mi-juillet, peu de temps restait pour envisager de recommencer tout le traitement bioinformatique avec de nouveaux outils non maîtrisés. Cela a néanmoins des inconvénients car Excel utilise de la mémoire vive et que ce logiciel est payant. De plus, il n'est pas utilisable sur un serveur de calcul qui utilise généralement UNIX.

De plus, il aurait été dommage de perdre le travail déjà effectué, surtout après avoir réussi à gérer des données avec de nombreux indels ce qui n'est pas le cas de tous les logiciels de mapping. De même, les données de sortie issues de Galaxy étaient propres et de bonne qualité car la localisation des indels et des SNP était bien gérée et cohérente avec nos données de polymorphisme.

Le programme mis au point, ajouté au contigage de reads pairés mappés ensemble, a l'avantage de gérer les indels et également d'éliminer les erreurs de séquençage. Or ces étapes permettront pour la suite de l'étude de ne pas compter des reads avec des erreurs de séquençage comme des haplotypes à part entière et donc d'augmenter la qualité du dosage. En effet, la poursuite de la réalisation du fichier Excel, concerne le dosage d'haplotypes. Or, nous avons vu que les rosiers ont des niveaux de ploïdies très différents. Ainsi, un rosier diploïde n'a que 2 allèles différents maximum et l'on ne devrait donc dénombrer pas plus de 2 haplotypes différents, 4 pour un tétraploïde... L'avantage de notre programme est qu'il sera adapté à ces différentes données. Ainsi, les résultats du dosage pourront être comparés aux résultats théoriques. Il pourra être envisagé de calculer un degré d'incertitude de notre dosage.

Le seuil d'erreur de 5% que nous avons fixé pourrait être modifié selon la ploïdie. De plus, les erreurs issues de la PCR n'ont pas été prises en compte dans notre analyse. Le manque d'informations sur les erreurs de la technologie Access Array que nous avons utilisé nous conduira à estimer cette erreur. Pour ce faire, il serait envisageable de calculer la quantité d'ADN utilisé dans la PCR à partir de la concentration de nos échantillons et de leur volume. A cette étape, une première incertitude sera liée à la méthode de dosage employé (Nanodrop®, fluorimétrie ou Qubit®) et dont les résultats sont parfois très différents selon les individus. A partir de ce résultat et de la taille du génome du rosier, il est possible d'en déduire, selon la ploïdie, un nombre moyen de copies d'allèles. On pourrait alors estimer le taux d'erreur en cas d'une erreur de PCR produite dès le premier cycle d'amplification.

Pour la suite de l'étude, il va également falloir prendre en compte la gestion du format d'entrée et les dispositions des données au sein du fichier .sam issus de Galaxy. Or, pour minimiser le nombre d'enregistrements manuels, les fichiers .sam seront issus d'une concaténation. Il sera donc nécessaire de récupérer les données de certaines colonnes et de les séparer par gène.

Enfin, il faut réaliser une automatisation du traitement de l'ensemble des fichiers Excel générés. En effet, le logiciel devra être adapté pour chacun des gènes et ne doit traiter qu'un individu à la fois. Il va donc falloir mettre au point un système de boucle permettant le traitement des fichiers Excel les uns après les autres.

4.5. Poursuite de la création des amorces du second lot de 48 amorces

Pour l'ensemble de l'étude, il reste à réaliser le séquençage de *DFR* et *FLS*, dont les amorces ont déjà été réalisées. De plus, il restera 27 couples d'amorces à mettre au point. Certains d'entre eux pourront servir à combler des manques de données notamment pour le dosage inter-séquences. Il est en plus prévu d'en utiliser au moins 20 pour séquencer des portions du génome choisies de façon aléatoire, afin de disposer de gènes a priori non sélectionnés et dont le polymorphisme sera comparé aux gènes candidats de l'étude (Lacombe, 2012). Il faudra néanmoins veiller à ce que le gène choisi ne fasse pas partie d'une famille multigénique et il pourrait être intéressant d'avoir des informations sur sa fonction. Les transcrits disponibles sur le site du transcriptome de Toulouse [24] seront utiles pour ce travail. Un seul couple d'amorces sera utilisé pour amplifier une partie seulement de ces gènes. Neuf de ces gènes ont d'ores et déjà été mis au point. Une fois ces 48 nouvelles amorces créées, la réalisation des 8 derniers runs pourra alors être lancée. Il est envisagé de réaliser une amplification de plusieurs couples d'amorces dans un même puits (multiplex). Cela permettra de mettre au point plus de couples d'amorces et donc d'amplifier et séquencer plus de gènes « aléatoires ». Des gènes de résistance, notamment à la tâche noire, pourraient être inclus dans l'étude.

4.6. Les analyses de diversité envisagées

A l'issue du traitement bioinformatique de l'ensemble des données, une analyse du polymorphisme et de la diversité nucléotidique sera réalisée. Celle-ci sera à adapter selon le succès de notre dosage d'haplotypes et du contigage inter-séquences. Les différentes études envisagées sont :

- Le calcul du polymorphisme nucléotidique qui permettra de comparer les groupes de rosiers en fonction leur variabilité pour les gènes-candidats
- Le calcul du F_{ST} ou indice de fixation qui servira à évaluer la différenciation des populations à partir des données de polymorphisme. En effet, plus cet indice est élevé et proche de 1, plus la différenciation est importante.
- L'étude de l'histoire de l'évolution des allèles des gènes candidats (augmentation ou diminution de la fréquence d'un allèle du gène au cours du temps) et comparaison avec les données des gènes « aléatoires », ce qui permettra de déterminer si un gène porte des traces de sélection (Lacombe, 2012).

Conclusion

Le travail réalisé au cours de ce stage a permis de mettre en place différentes stratégies afin d'effectuer par la suite une analyse optimisée de la diversité des gènes candidats.

La première stratégie mise au point concernait la création des amorces. Elle a servi à créer des couples qui amplifiaient de façon spécifique l'ensemble d'un gène tout en veillant à respecter des contraintes liées aux conditions de fonctionnement de la PCR Access Array et à la conservation de zones de chevauchement intra et inter-séquences. Les tailles de celles-ci ont été discutées et il serait conseillé de les augmenter afin d'obtenir des haplotypes plus complets à l'issue du contigage entre deux reads et également faciliter la phase de contigage inter-séquences. Cette dernière étape n'a pas été réalisée dans le cadre de cette étude mais en sera une poursuite directe. Enfin, il s'est avéré que certains couples ont mal ou peu amplifié les portions d'ADN souhaitées et il pourra être envisagé d'en changer certains ou d'en ajouter.

La stratégie de traitement des données sur l'instance Galaxy a permis d'obtenir des reads mappés propres et de bonne qualité. La réflexion des paramètres à fixer pour le quality trimming a servi à réaliser un compromis entre la conservation des bases de mauvaise qualité et la perte d'informations due à la suppression d'un nombre important de bases. C'est donc une méthode de fenêtre glissante qui s'est révélée être la mieux adaptée à nos objectifs. La sélection du logiciel de mapping a été choisie et raisonné en fonction de nos données de séquençage tout en prenant soin de vérifier la bonne gestion des indels et la position des SNP. Cependant, un problème majeur pour la suite de l'étude reste un manque d'automatisation de l'instance Galaxy.

Enfin, il a fallu s'adapter à une incompatibilité de script et mettre au point un code permettant de réaliser le contigage des reads, le filtrage des erreurs de séquençage et le dosage des haplotypes. Ce programme a permis de prendre en compte des caractéristiques propres au rosier telles que les différentes ploïdies des individus et le grand nombre d'indels au sein des séquences. Cependant, le code doit être adapté pour chacun des gènes candidat et le temps de calcul est plus élevé que sur un serveur. Néanmoins, le programme est en cours d'automatisation, ce qui permettra un gain de temps et une limitation des manipulations par l'utilisateur.

***

Les résultats obtenus à la fin de ce stage vont être utilisés afin d'obtenir les jeux de données traités qui seront utilisés pour l'analyse de la diversité des gènes candidats. Celle-ci permettra de comparer les groupes de rosiers entre eux et d'observer les différences entre les individus mais aussi de déterminer l'histoire de l'évolution des allèles de ces gènes.

Ils seront également utilisés afin de réaliser une comparaison entre les données phénotypiques dont dispose l'équipe GDO et les différents génotypes qui auront été analysés. Cela permettra d'identifier si certains génotypes expliquent des phénotypes observés et donc si les gènes candidats sélectionnés étaient pertinents.

A terme, ce travail s'intègrera dans le projet FloRHiGe et aidera en la mise en place d'une meilleure gestion de la diversité génétique et de sa conservation par la création d'une core-collection. Celle-ci servira à raisonner la conservation des différentes espèces de rosiers en minimisant le nombre d'individus à garder tout en maximisant la diversité génétique qu'ils représentent.

Bibliographie

Aguilar-Martínez J.A., Poza-Carrión C., Cubas P. (2007). Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *The Plant Cell*, *19*, 2, pp.458-472.

Altschul S., Gish W., Miller W., Myers E., Lipman D. (1990). *Journal of Molecular Biology*, 215, 3, pp.403–410.

Blankenberg D., Von Kuster G., Coraor N., Ananda G., Lazarus R., Mangan M., Nekrutenko A., Taylor J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, 19, 10, pp.1-21.

Cairns T. (2003). Horticultural Classification Schemes. In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.117-124.

Crane Y. M., Byrne D. H. (2003). Karyology. In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.267-273.

De Vries D.P., Van Keulen H.A., De Bruyn J.W. (1974). Breeding research on rose pigments : The occurrence of flavonoids and carotenoids in rose petals. *Euphytica*, 23, 2, pp.447-457.

Debener T., Janakiram T., Mattiesch L. (2000). Sports and seedlings of rose varieties analysed with molecular markers. *Plant Breeding*, 119, 1, pp.71-74.

Forkmann G. (2003). Flavonoid molecular Biology. In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.256-263.

Foucher F., Chevalier M., Corre C., Soufflet-Freslon V., Legeai F., Hibrand-Saint Oyant L. (2008). New resources for studying the rose flowering process. *Genome*, 51, 10, pp.827-837.

Fukuchi-Mizutani M., Akagi M., Ishiguro K., Katsumoto Y., Fukui Y., Togami J., Nakamura N., Tanaka Y. (2011). Biochemical and molecular characterization of anthocyanidin/flavonol 3-glucosylation pathways in Rosa x hybrid. *Plant Biotechnology*, 28, pp.239-244.

Gallais A., Bannerot H. (1992). Amélioration des espèces végétales cultivées : Objectifs et critères de selection. *Mieux Comprendre*. Paris, INRA Editions, 768p.

Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J., Miller W., Kent W.J., Nekrutenko A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research,* 15, 10, pp.1451-1455.

Gudin, S. (2000). Rose: genetics and breeding. *Plant breeding reviews*, 17, pp.159-190.

Guoliang W. (2003). History of roses in cultivation/ancient Chinese roses. In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.385-395.

Haudebourg M.T. (1998). *Roses et jardins*. Hachette Pratique, 477p.

Holton T.A., Cornish E.C. (1995). Genetics and biochemistry of anthocyanin biosynthesis. *The Plant Cell*, 7, pp.1071-1083.

Iwata H., Gaston A., Remay A., Thouroude T., Jeauffre J., Kawamura K., Hibrand-Saint Oyant L., Araki T., Denoyes B., Foucher, F. (2012). The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry. *The Plant Journal*, 29, 1, pp.116-125.

Jay M., Biolley J-P., Fiasson J-L., Fiasson K., Gonnet J-F., Grossi C., Raymond O., Viricel M-R. (2003). Anthocyanins and other flavonoid Pigments. In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.248-255.

Jian H., Zhang H., Tang K., Li S., Wang Q., Zhang T. Qiu X., Yan H. (2010). Decaploidy in *Rosa praelucens* Byhouwer (Rosaceae) endemic to zhongdian plateau, Yunnan, China. *Caryologia*, 63, 2, pp.162-167.

Johnson E.T., Ryu S., Yi H., Shin B., Cheong H., Choi G. (2001). Alteration of a single amino acid changes the substrate specificity of dihydroflavonol 4-reductase. *The Plant Journal*, 25, 3, pp.325-333

Joyaux F., Lévêque G. (2001). *La rose, une passion française (1778-1914)*. Editions Complexe. 249 p.

Katsumoto Y., Fukuchi-Mizutani M., Fukui Y., Brugliera F., Holton T.A., Karan M., Nakamura N., Yonekura-Sakakibara K., Togami J., Pigeaire A., Tao G-Q., Nehra N.S., Lu C-Y., Dyson B.K., Tsuda S., Ashikari T., Kusumi T., Mason J.G., Tanaka Y. (2007). Engineering of the rose flavonoid biosynthetic pathway successfully generated blue-hued flowers accumulating delphinidin. *Plant Cell Physiology*, 48, 11, pp.1589-1600.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Mentjies P., Drummond A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*, 12, pp.1647-1649.

Kobayashi Y., Kaya H., Goto K., Iwabuchi M., Araki T. (1999). A pair of related genes with antagonistic roles in mediating flowering signals. *Science*, 286, 5446, pp.1960-1962.

Krussmann G. (1981). The history of the modern garden rose. *The Complete Book of Roses*, pp.67-105.

Lacombe T. (2012). Contribution à l'étude de l'histoire évolutive de la vigne cultivée (Vitis vinifera L.) par l'analyse de la diversité génétique neutre et de gènes d'intérêt. Doctorat Sciences agronomiques, Spécialité Evolution, Ecologie, ressources génétiques et Paléontologie, Montpellier SupAgro, Montpellier, 328 p.

Machenaud J. (2010). Biosynthèse du 2-phényléthanol et sécrétion du parfum chez la rose. Doctorat Biologie et Physiologie végétale, Université Jean Monet, Saint-Etienne, 184 p.

Magnard J.L., Roccia A., Caissard J.C., Vergne P., Sun P., Hecquet, R., Dubois A., Hibrand-Saint Oyant L., Jullien F., Nicolè F., Raymond O., Huguet S., Baltenweck R., Meyer, S., Claudel P., Jeauffre J., Rohmer M., Foucher F., Hugueney P., Bendahmane M., Baudino S. (2015). Biosynthesis of monoterpene scent compounds in roses. *Science*, 349, 6243, pp.81-83.

Maia N., Vénard P. (1976). Cytotaxonomie du genre Rosa et origine des rosiers cultivés – Travaux sur le rosier de serre menés au C.R.A. d'Antibes. *Fédération Nationale des Producteurs de l'Horticulture et des Pépinières*, INRA Editions, Paris, pp.7-20.

Marriott M. (2003). Modern (Post-1800). In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.402-409.

Medina N., Broka A., Lacey S., Lin H., Klings E.S., Baldwin C.T., Steinberg M.H., Sebastiani P. (2012). Comparing Bowtie and BWA to Align Short Reads from a RNA-Seq Experiment. *6th International Conference on Practical Applications of Computational Biology & Bioinformatics Advances in Intelligent and Soft Computing*, 154, pp.197-207.

Meynet J. (2001). Les rosiers cultivés, une très longue histoire d'exploitation de la biodiversité seulement pour le plaisir et l'art de vivre. *Les dossiers de l'environnement de l'INRA*, 21, pp.113-118.

Ogata J., Kanno Y., Itoh Y., Tsugawa H., Suzuki M. (2005). Anthocyanin biosynthesis in roses. *Nature Publishing group*, 435, pp.757-758.

Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Roccia A. (2013). Etude de deux gènes impliqués dans la biosynthèse du parfum chez le genre Rosa *L*. (Rosaceae). Doctorat Biologie et Physiologie végétale, Université Jean Monet, Saint-Etienne, 200 p.

Rozen S., Skaletsky H. (2000). *Primer3 on the WWW for general users and for biologist programmers.* In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Krawetz S., Misener S. eds, Humana Press, Totowa, pp.365-386.

Schulz H. (2003). Odoriferous Substances and Pigments. In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.231-240.

Suzuki K-I., Tsuda S., Y. Fukui, Fukuchi-Mizutani M., Yonekura-Sakakibara K., Tanaka Y., Kusumi T. (2014). Molecular characterization of rose flavonoid biosynthesis genes and their application in Petunia, *Biotechnology & Biotechnological equipment*, 14, 2, pp.56-62.

Swerdlow H.P., Gu Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13, 1, pp.341.

Tanaka Y., Brugliera F. (2006). Flower colour and cytochromes P450. *Phytochemistry Reviews*, 2006, vol. 5, no 2-3, p. 283-291.

Tanaka Y., Fukui Y., Fukuchi-Mizutani M., Holton T.A., Higgins E.,Kusumi T. (1995). Molecular cloning and characterization of Rosa hybridadihydroflavonol 4-reductase gene. *Plant and cell physiology*, 36, 6, pp.1023-1031.

Testu C. (1984). *Les roses anciennes*. La Maison rustique, 247p.

Vukosavljev M., Zhang J., Esselink G.D., van't Westende W.P.C., Cox P., Visser R.G.F., Arens P., Smulders M.J.M. (2013). Genetic diversity and differentiation in roses: A garden rose perspective. *ScientiaHorticulturae*, 162, pp.320-332.

Wissemann V. (2003). Conventional taxonomy (wild roses). In: *Encyclopedia of Rose Science, Volume 1*. Roberts A.V., Debener T., Gudin S. eds, Elsevier, Oxford, pp.111-117.

Zhao D., Tao J. (2015). Recent advances on the development and regulation of flower color in ornamental plants. *Frontiers in Plant Science*, 6, pp.261-313.

Zhu C., Bai C., Sanahuja G., Yuan D., Farré G., Naqvi S., Shi L., Capell T., Christou, P. (2010). The regulation of carotenoid pigmentation in flowers. *Archives of Biochemistry and Biophysics*, 504, 1, pp.132-141.

Sitographie

[1] FranceAgriMer (2015). Données et bilans : Végétaux d'ornement, achats des ménages en 2014. http://www.franceagrimer.fr/content/download/38833/358643/file/BIL-HOR%20Achats%20des%20m%C3%A9nages%20en%202014.pdf (consulté le 10/08/2015)
[2] FranceAgriMer (2014). La rose en 2013-2014 : bilan de campagne. https://www.rnm.franceagrimer.fr/bilan/rose_rnm.pdf (consulté le 10/08/15)
[3] Société française des roses (2015). Résultats du 85ème Concours International de Roses Nouvelles de Lyon. http://societefrancaisedesroses.asso.fr/fr/actualites/concours.htm (consulté le 10/08/2015)
[4] Julien Jeauffre (2013). Les équipes de l'unité de recherche. https://www6.angers-nantes.inra.fr/irhs/Recherche (consulté le 10/08/2015)
[5] Université Jean Monet Saint-Etienne (2015). Laboratoire de Biotechnologies Végétales appliquées aux Plantes Aromatiques et Médicinales http://portail.univ-st-etienne.fr/bienvenue/recherche/laboratoire-de-biotechnologies-vegetales-appliquees-aux-plantes-aromatiques-et-medicinales-26130.kjsp (consulté le 10/08/2015)
[6] Laboratoire de Reproduction et Développement des Plantes (2015). Morphogenèse florale. http://www.ens-lyon.fr/RDP/spip.php?rubrique23&lang=fr (consulté le 10/08/2015)
[7] Nathalie Mansion (2015). Présentation du projet. http://www6.inra.fr/florhige/Presentation-du-projet (consulté le 08/04/2015)
[8] Wikipédia (2015). 'Old Blush'. https://en.wikipedia.org/wiki/Rosa_'Old_Blush' (consulté le 15/06/2015)
[9] Société française des roses (2015). Choisir son rosier. http://societefrancaisedesroses.asso.fr/fr/rosiers_et_roses/choisir_rosier.htm (consulté le 15/06/2015)
[10] Guillot (2015). Rosier 'La France'. http://www.roses-guillot.com/rosiers-718/327-rosier-la-france.html (consulté le 15/06/2015)
[11] Wikipédia (2015). *Rosa foetida*. https://en.wikipedia.org/wiki/Rosa_foetida (consulté le 15/06/2015)
[12] Rhode Island Rose Society (2004). Rosier 'Max Graf'. http://www.rirs.org/cranford2004.htm (consulté le 15/06/2015)
[13] CoGEPEDIA (2015). Sequenced plant genomes. https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes (consulté le 23/06/2015)
[14] Plant and Animal genome (2012). The Rose Genome Sequence Initiative. https://pag.confex.com/pag/xx/webprogram/Paper4563.html (consulté le 23/06/2015)
[15] Jean-Claude Caissard et al. (2013). Le séquençage du génome du rosier : pourquoi faire ? http://www.jardinsdefrance.org/le-sequencage-du-genome-du-rosier-pourquoi-faire/ (consulté le 28/05/2015)
[16] Vincent Charbonnier (2015). Le génome de la rose en cours de décryptage http://www.lesechos.fr/04/05/2015/lesechos.fr/02146862059_le-genome-de-la-rose-en-cours-de-decryptage.htm#vzXsmDbXFfsfOuol.99 (consulté le 28/05/2015)
[17] Sierraflowerfinder (2015). Rose 'Super Green'. http://www.sierraflowerfinder.com/en/d/super-green/4673 (consulté le 22/06/2015)
[18] Dubois L.A.M. (1980). Pigments and Petal Colors. http://bulbnrose.x10.mx/Roses/breeding/Dubois/Dubois.html (consulté le 16/04/2015)
[19] Geneious (2015). Page d'accueil. http://www.geneious.com (consulté le 03/04/2015)
[20] Goecks J., Nekrutenko A., Taylor J. and The Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.

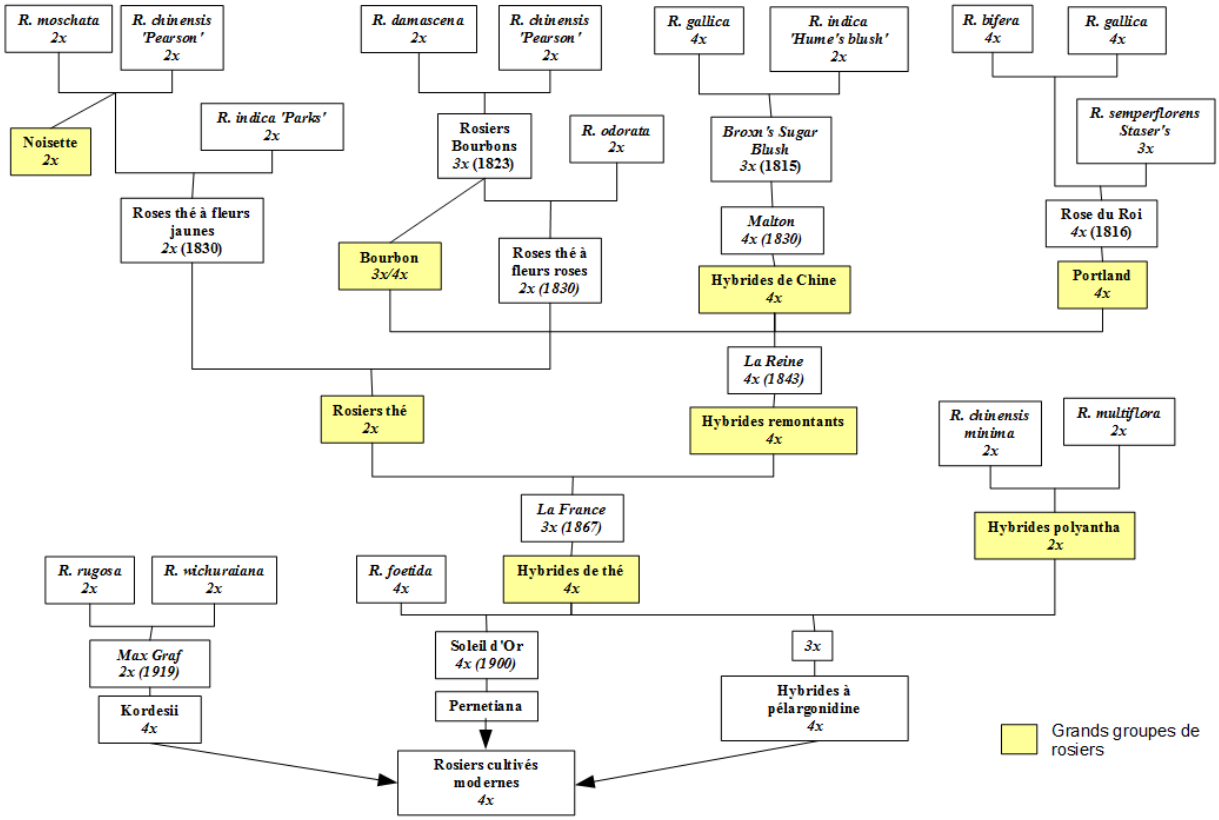
http://www.biomedcentral.com/content/pdf/gb-2010-11-8-r86.pdf *(consulté le 03/08/2015)*
[21] NCBI (2015). Welcome to NCBI.
http://www.ncbi.nlm.nih.gov/ (consulté le 06/04/2015)
[22] TAIR (2015). The Arabidopsis Information Resource.
https://www.arabidopsis.org/ (consulté le 06/04/2015)
[23] GDR (2015). Welcome to the Genome Database for Rosaceae
http://www.rosaceae.org/ (consulté le 16/06/2015)
[24] INRA Toulouse (2015). *Rosa chinensis* Genomics.
https://lipm-browsers.toulouse.inra.fr/plants/R.chinensis (consulté le 16/06/2015)
[25] MillenniumScience (2011) Fluidigm Dynamic Array.
https://www.youtube.com/watch?v=s9HUhuCbbhU (consulté le 16/07/2015)
[26] Illumina (2015). Access Array for Illumina Sequencing Systems User Guide. https://www.fluidigm.com/documents (consulté le 16/07/2015)
[27] Illumina (2015). Illumina reference.
http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq-system-user-guide-15027617-l.pdf (consulté le 16/07/2015)
[28] Genomic Services Lab at HudsonAlpha (2015). Platforms.
http://gsl.hudsonalpha.org/information/platforms/sequencing (consulté le 05/08/2015)
[29] Rutgers School of environnemental and Biological Science (2015). Illumina MiSeq Genome Analyser.
http://dblab.rutgers.edu/home/html/genome_analyzer.php (consulté le 06/08/2015)
[30] Illumina Incorporation (2013). Illumina Technology Sequencing.
https://www.youtube.com/watch?v=womKfikWlxM (consulté le 16/07/2015)
[31] Illumina Platform (2015). New generation sequencing method.
https://www.youtube.com/watch?v=Zqr8_KiuzHU (consulté le 16/07/2015)
[32] Geneious (2012). Geneious support.
https://support.geneious.com/entries/22121098-What-s-the-difference-between-alignment-de-novo-assembly-and-map-to-reference- (consulté le 03/08/2015)
[33] Blankenberg D., Gordon A.., Von Kuster G, Coraor N., Taylor J., Nekrutenko A. and Galaxy Team. (2010) Manipulation of FASTQ data with Galaxy.
http://bioinformatics.oxfordjournals.org/content/26/14/1783.short (consulté le 23/07/2015)
[34] Langmead B., Trapnell C., Pop M. and Salzberg S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.
http://www.biomedcentral.com/content/pdf/gb-2009-10-3-r25.pd (consulté le 03/08/2015)
[35] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. (2009). The Sequence Alignment/Map format and SAMtools.
http://bioinformatics.oxfordjournals.org/content/25/16/2078.short (consulté le 24/07/2015)
[36] Equipe bioinformatique Roscoff (2013). Le contrôle qualité sur les données fastq.
http://biow.sb-roscoff.fr/ecole_bioinfo/training_material/snp/exome_qc_OInizan.pdf (consulté le 13/05/2015)
[37] Chen C., Khaleel S.S., Huang H., Wu C.H. (2014). Software for pre-processing Illumina next-generation sequencing short read sequences.
http://www.biomedcentral.com/content/pdf/1751-0473-9-8.pdf (consulté le 24/07/2015)
[38] Illumina (2015). System specification Sheet : Sequencing MiSeq® System.
http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf (consulté le 20/05/2015)
[39] SEQanswers (2009). Insert size and Fragment size
http://seqanswers.com/forums/showthread.php?t=15511 (consulté le 20/05/2015)

[40] Li H., Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler

Sitographie

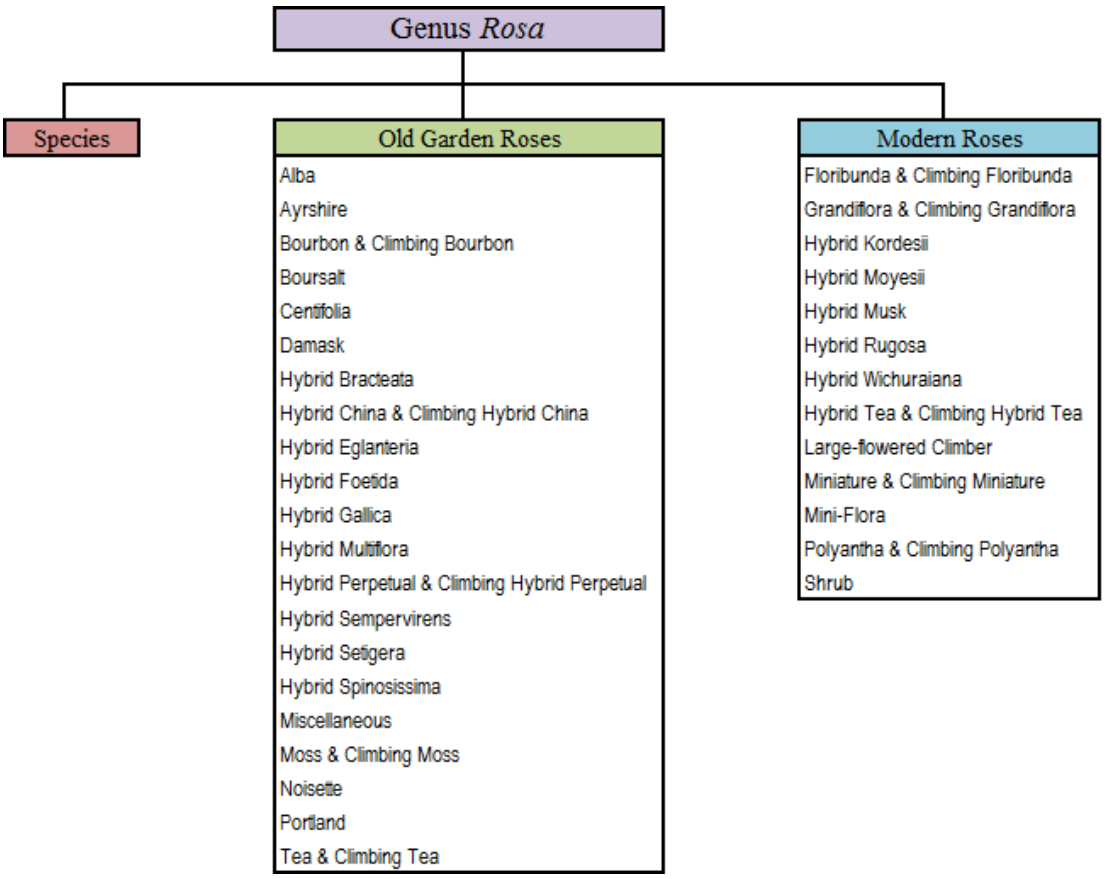
Transform.
http://www.ncbi.nlm.nih.gov/pubmed/19451168 (consulté le 19/06/2015)
[41] Choudhury O., Bernhard J. (2014). Performance Analysis of Genome Alignment Tools: BWA and Bowtie.
http://netscale.cse.nd.edu/cms/pub/Edu/GradOSF12InterimDraft/Bernhard-Choudhury_InterimReport.pdf (consulté le 23/06/2015)
[42] Johns Hopkins University (2015). Bowtie2, fast and sensitive read alignment.
http://bowtie-bio.sourceforge.net/bowtie2/faq.shtml (consulté le 07/07/2015).
[43] Johns Hopkins University (2015). Bowtie, an ultrafast memory-efficient short read aligner.
http://bowtie-bio.sourceforge.net/news.shtml (consulté le 23/06/2015)
[44] Boekhoff S. (2015). Fasta DNA codes.
http://www.boekhoff.info/?pid=data&dat=fasta-codes (consulté le 24/07/2015)
[45] Biostars (2014). Bowtie2 parameters for best alignment.
https://www.biostars.org/p/103705/ (consulté le 24/07/2015)
[46] Giannoulatou E., Park S.H., Humphreys, D.T., Ho J.W. (2014). Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie
http://www.biomedcentral.com/1471-2105/15/S16/S15 (consulté le 25/06/2015)
[47] The SAM/BAM Format Specification Working Group (2015). Sequence Alignment/Map Format Specification.
https://samtools.github.io/hts-specs/SAMv1.pdf (consulté le 25/06/2015)
[48] Pico (2010). SAMtool bitwise flag meaning explained: how to understand samflags without pains.
https://ppotato.wordpress.com/2010/08/25/samtool-bitwise-flag-paired-reads/ (consulté le 24/07/2015)
[49] Bordinstitute (2015). This utility explains SAM flags in plain English.
http://broadinstitute.github.io/picard/explain-flags.html (consulté le 23/06/2015)
[50] Biostars (2014). Perplexing Bowtie 2 results
https://www.biostars.org/p/120699/ (consulté le 23/06/2015)
[51] Biostars (2014). How does bowtie2 assign MAPQ scores?.
https://www.biostars.org/p/110958/ (consulté le 01/07/2015)
[52] John Urban (2014). How does bowtie2 assign MAPQ scores?
http://biofinysics.blogspot.fr/2014/05/how-does-bowtie2-assign-mapq-scores.html (consulté le 01/07/2015)
[53] Biostars (2013). Illumina MiseqPhix Quality Control (Sequencing Error Rates).
https://www.biostars.org/p/73346/ (consulté le 06/08/2015)
[54] CLC bio, a QIAGEN® compagny (2014). Tutorial, Read Mapping in Detail.
http://www.clcbio.com/files/tutorials/Read_mapping_in_detail.pdf (consulté le 09/07/2015)

Annexe I : Histoire simplifiée du rosier



Histoire simplifiée du rosier d'après Maïa et Vénard, 1976 et Meynet, 2001.

Annexe II : Classification horticole de l'American Rose Society



Classification horticole de l'American Rose Society d'après Cairns 2003.

Annexe III : Dosage d'ADN par fluorimétrie au Hoescht (données partielles car le fichier de calcul du phage Lambda n'est pas communiqué)

Annexe III : Dosage d'ADN par fluorimétrie au Hoescht (données partielles car le fichier de calcul du phage Lambda n'est pas communiqué)

Matériel :
Fluorimètre : TECAN Infinite ® 200
Plaques noires : 96well nunc maxisorb black réf

Réactifs spécifiques :
Hoescht33258 : Bis-benzimide - Sigma Aldrich / Réf pour 25mg - excitation : 365nm/émission : 450nm
Solutions :
Tris 1M pH8
EDTA 0,5M pH8

Méthodologie :
Préparer la solution de TE à partir de Tris 1M pH8 et EDTA 0.5M pH8

Quantité pour	1 plaque	2 plaques	3 plaques	4 plaques
Tris 1M pH8	0.5 ml	1 ml	1.5 ml	2 ml
EDTA 0.5M pH8	0.1 ml	0.2 ml	0.3 ml	0.4 ml
H2O	49.4 ml	98.8 ml	148.2 ml	197.6 ml
Volume final	50 ml	100 ml	150 ml	200 ml

Préparer une solution de Hoescht à 1mg/ml à partir de la solution mère à 25 mg/ml (4µl dans 96µl).
Préparer la solution de TE + Hoescht (sera utilisée en dilution finale au ½)

Quantité pour	**1 plaque**	**2 plaques**	**3 plaques**	**4 plaques**
TE	12 ml	24 ml	36ml	48 ml
Hoescht à 1 mg/ml	2,4 µl	4.8 µl	7 µl	9.6 µl
Volume final	12 ml	25 ml	37 ml	50 ml

Dans la plaque noire, distribuer 100µL TE-Hoechst dans chaque puits (pour la gamme et pour les dosages)
Dans la plaque noire, distribuer 100µL de solution TE-Hoechst dans les puits qui contiendront le blanc (TE), les différents points de dilution de la gamme, ainsi que ceux qui contiendront les échantillons d'ADNg à doser.

Préparation de la gamme du phage Lambda (standard-fichier Excel de calcul non fourni)

Préparation des échantillons d'ADN à doser
Faire une dilution au 1/80ème :
dilution au 1/40ème : 5µl de la solution mère d'ADN + 195 µl de TE
dilution au ½ : 100µl de la solution au 1/40ème + 100 µ de TE+Hoescht (déjà distribué dans la plaque noire)
Réaliser une dilution au 1/40ème : 5µl de la solution mère d'ADN + 195 µl de TE.
Ajouter 100 µl de cette dilution dans chaque puits de la plaque noire sauf colonne 1 qui correspond à la gamme

Annexe IV : Liste des séquences sources utilisées

Gène	Espèce	Nom de la séquence	Base	Nature	Longueur (pb)	ID
DFR	*Fragaria vesca*	DFR bifunctional dihydroflavonol 4-reductase/flavanone 4-reductase [*Fragaria vesca* (wild strawberry)]	NCBI Genbank	gene	2239	NC_020492.1
				ARNm	1359	
				cds	1050	
	Rosa chinensis	Rosa chinensis dihydroflavonol 4-reductase (DFR) mRNA, complete cds	NCBI Genbank	gene	1232	KF734592.1
				cds	1050	
	Rosa hybrida	Rosa hybrida mRNA for dihydroflavonol 4-reductase, complete cds	NCBI Genbank	cds	1050	D85102.1
	Rosa hybrid 'Noblesse'	Rosa hybrid cultivar RhDFR mRNA for dihydroflavonol 4-reductase, complete cds, cultivar: Noblesse	NCBI Genbank	cds	1050	AB490072.1
F3'H	*Fragaria vesca*	Fragaria vesca subsp. vesca linkage group LG5, FraVesHawaii_1.0, whole genome shotgun sequence	NCBI Genbank	gene	3140	NC_020495.1
				ARNm	1539	
				cds	1539	
	Fragaria x ananassa	Fragaria x ananassa F3'H mRNA for flavonoid 3'-hydroxylase, complete cds	NCBI Genbank	gène	1775	AB665441.1
				cds	1533	
FLS	*Fragaria x ananassa*	Fragaria x ananassa flavonol synthase (FLS) mRNA, complete cds	NCBI Genbank	promoteur et 5'UTR	1211	DQ087252.1
				cds	1008	
				ARNM partiel	1008	
	Rosa hybrid 'Kardinal'	Rosa hybrid cultivar 'Kardinal' FLS mRNA for flavonol synthase, complete cds	NCBI Genbank	gène	1399	AB038247.1
				cds	1008	
	Rosa multiflora 'Duohua'	Rosa multiflora cultivar Duohua flavonol synthase mRNA, complete cds	NCBI Genbank	cds	1008	KP090455.1
	Rosa rugosa 'Fenghua'	Rosa rugosa cultivar Fenghua flavonol synthase (FLS) mRNA, complete cds	NCBI Genbank	gène partiel	1008	KM099095.1
				cds	1008	
ZDS	*Fragaria vesca*	PREDICTED: Fragaria vesca subsp. vesca zeta-carotene desaturase, chloroplastic/chromoplastic (LOC101296114), mRNA	NCBI Genbank	gène	4558	XM_004301977.2
				ARNm	2241	
				cds	1710	
	Fragaria x ananassa	Fragaria x ananassa zeta-carotene desaturase protein (zds) mRNA, complete cds	NCBI Genbank	cds	1710	FJ795343
				gène partiel	2148	
AGAMOUS	*Fragaria vesca*	LOC101303096 floral homeotic protein AGAMOUS [*Fragaria vesca* (wild strawberry)]	NCBI Genbank	gene	5914	NC_020493.1
				cds	747 et 756	
				ARNm	1236 et 930	
	Rosa hybrida	Rosa hybrida AGAMOUS protein (RAG) mRNA, complete cds	NCBI Genbank	gene	1116	U43372.1
				cds	747	

Annexe IV : Liste des séquences sources utilisées

Gène	Espèce	Nom de la séquence	Base	Nature	Longueur (pb)
FT	'Old Blush'	RoFT1_ADNg_scf3636	Equipe GDO	gène	1315
		RoFT1_ADNg_scf_4318	Equipe GDO	gène	1238
KSN	*Rosa chinensis*	KSN_Rosa chinensis spontanea	Equipe GDO	gène	1184
	Rosa chinensis	KSN_Rosa OB_Transposon	Equipe GDO	gène	10114
	'Old Blush'	RoKSN_cDNA_OB	Equipe GDO	cds	519
NUDX	'Old Blush'	RcNUDX1-OB1a (ATG-STOP)	Equipe de Saint-Etienne	gène	622
	'Old Blush'	RhNUDX1-OB1b (ATG-STOP)	Equipe de Saint-Etienne	gène	618
	Papa Meilland'	RhNUX1-PM1 (ATG-STOP)	Equipe de Saint-Etienne	gène	618
	Rosa wichurana	RwNUDX1-RW1 (ATG-STOP)	Equipe de Saint-Etienne	gène	618
	Rosa wichurana	RwNUDX1-RW2 (ATG-STOP)	Equipe de Saint-Etienne	gène	759
	'Old Blush'	RcNUDX1-OB2a (ATG-STOP)	Equipe de Saint-Etienne	gène	729
	'Old Blush'	RhNUDX1-OB2b-(ATG-STOP)	Equipe de Saint-Etienne	gène	766
	'Old Blush'	RhNUDX1-OB2c-(ATG-STOP)	Equipe de Saint-Etienne	gène	799
PAAS	*Rosa hybrid* 'Fragrant Cloud'	DQ192639	Equipe de Saint-Etienne	gène	1775
			Equipe de Saint-Etienne	cds	1527
	'Old Blush'	PAAS scaffold_1004	Equipe de Saint-Etienne	gène	1527
	'Old Blush'	PAAS scaffold_1103	Equipe de Saint-Etienne	gène	1527
	Rosa wichurana	RhPAAS a1	Equipe de Saint-Etienne	gène	1527
	Rosa wichurana	RhPAAS a2	Equipe de Saint-Etienne	gène	1527
	H190'	RhPAAS a3	Equipe de Saint-Etienne	gène	1527
BRC1	*Rosa hybrid*	Rosa hybrid cultivar BRC1 protein (BRC1) mRNA, complete cds	GDR	cds	1386
	Fragaria vesca	LG5:8361961..8366960	GDR	gène	1486
				ARNm	1486
				cds	1278

Annexe V : Protocole fourni par Qiagen® : Quick-StartProtocol DNeasy® 96 Plant Kit

Quick-Start Protocol

DNeasy® 96 Plant Kit

The DNeasy 96 Plant Kit (cat. no. 69181) can be stored at room temperature (15–25°C) for up to 1 year.

For more information, please refer to the *DNeasy Plant Handbook* and the *TissueLyser Handbook*, which can be found at www.qiagen.com/handbooks.

For technical assistance, please call toll-free 00800-22-44-6000, or find regional phone numbers at www.qiagen.com/contact.

Notes before starting

- This protocol is for purifying DNA from 2 x 96 samples of fresh plant tissue.
- Ensure that you are familiar with operating the TissueLyser and the QIAGEN® 96-Well-Plate Centrifugation System.
- Perform all centrifugation steps at room temperature (15–25°C).
- If necessary, redissolve any precipitates in Buffer AP1 and Buffer AW1 concentrates.
- Add ethanol to Buffer AW1 and Buffer AW2 concentrates.
- Preheat Buffer AP1 to 65°C.
- Prepare a fresh working lysis solution: For 2 x 96 samples, combine 90 ml Buffer AP1, 225 µl RNase A, and 225 µl Reagent DX.

1. Place up to 50 mg leaves into each tube in 2 collection microtube racks.
2. Add 1 tungsten carbide bead to each collection microtube.
3. Pipet 400 µl working lysis solution into each collection microtube. Tightly seal the microtubes using the caps provided.
4. Assemble each rack of collection microtubes into the TissueLyser.
5. Grind the sample for 1.5 minutes at 30 Hz.

April 2012

QIAGEN®

Sample & Assay Technologies

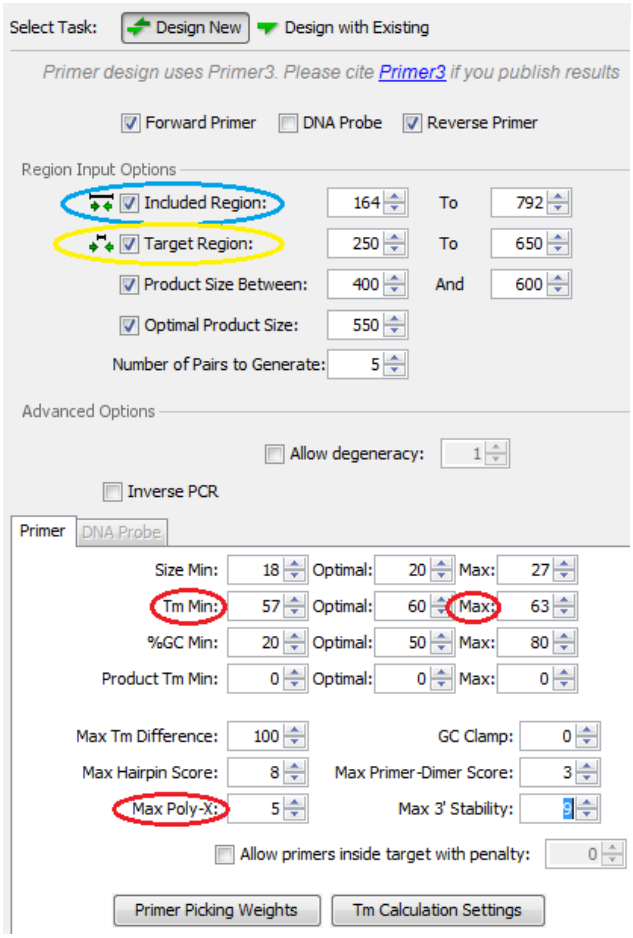
6. Reassemble the racks so that the collection microtubes nearest the TissueLyser in steps 4 and 5 are now furthest from the TissueLyser.
7. Grind the samples for another 1.5 min at 30 Hz.
8. Centrifuge to collect any solution from the caps.
9. Add 130 µl Buffer P3 to each collection microtube and reseal using new caps.
10. Place a clear cover over each rack and shake vigorously up and down for 15 s. Centrifuge to collect any solution from the caps.
11. Incubate the collection-microtube racks for 10 min at –20°C.
12. Centrifuge the collection-microtube racks for 5 min at 3800 x g (6000 rpm).
13. Transfer 400 µl of each supernatant to a new collection microtube.
14. Add 600 µl of Buffer AW1 to each sample. Close microtubes with new caps.
15. Place a clear cover over each rack and shake vigorously up and down for 15 s. Centrifuge to collect any solution from the caps.
16. Place 2 DNeasy 96 plates on top of S-Blocks. Mark the DNeasy 96 plates for later sample identification.
17. Transfer 1 ml of each sample to each well of the DNeasy 96 plates.
18. Seal each DNeasy 96 plate with an AirPore Tape Sheet. Centrifuge for 4 min at 3800 x g. If lysate remains in the DNeasy 96 plates after centrifugation, centrifuge for another 4 min.
19. Remove the tape. Add 800 µl Buffer AW2 to each sample.
20. Centrifuge for 15 min at 3800 x g without tape to dry the membranes.
21. Place each DNeasy 96 plate on a new Elution Microtubes RS rack.
22. Add 100 µl Buffer AE and seal with new AirPore Tape Sheets. Incubate for 1 min at room temperature (15–25°C). Centrifuge for 2 min at 3800 x g.
23. Repeat step 22. Seal the Elution Microtubes RS with new caps to store DNA.

Sample & Assay Technologies

Annexe VI : Températures et durées des cycles d'amplification de la PCR Access Array de Fluidigm® [26]

PCR Stages	Number of Cycles
50°C 2 minutes	1
70°C 20 minutes	1
95°C 10 minutes	1
95°C 15 seconds 60°C 30 seconds 72°C 1 minute	10
95°C 15 seconds 80°C 30 seconds 60°C 30 seconds 72°C 1 minute	2
95°C 15 seconds 60°C 30 seconds 72° 1 minute	8
95°C 15 seconds 80°C 30 seconds 60°C 30 seconds 72°C 1 minute	2
95°C 15 seconds 60°C 30 seconds 72°C 1 minute	8
95°C 15 seconds 80°C 30 seconds 60°C 30 seconds 72°C 1 minute	5

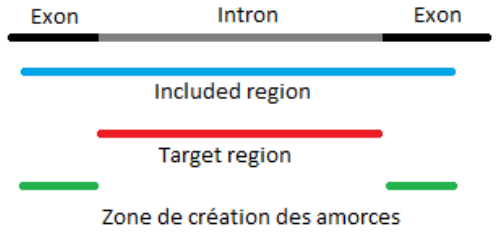
Annexe VII : Paramètres par défaut imposés pour la création des couples d'amorces sur Primer 3



Remarque : Distinction entre Included region (entourée en bleu) et Target region (entourée en jaune) :

- Included region = zone comprenant l'ensemble du fragment de la séquence que l'on souhaite amplifier plus la zone où se fixeront les amorces (bleu).
- Target region = zone que l'on souhaite amplifier mais où les amorces ne se fixeront pas (vert).

Cette distinction permet de placer les amorces de façon optimale en évitant les SNP, des introns…



2015, N. Jacquier

Annexe VIII : Tableau récapitulatif contenant l'ensemble des informations de chacune des amorces mises au point

Nom du couple d'amorce	Taille de l'amplicon	Amorce Forward				Amoce Reverse			
		Séquence	pb	Tm	Information	Séquence	pb	Tm	Information
FT_E1E2	348	TGTTGTCGGGAGAGTGATCGGAGA	24	58,9	1 (Tm inférieure + 1 SNP)	TCACTTGGACTGGGTGCATCAGG	23	58,57	1 (Tm inférieure)
FT_E1E4	741	ACCTCGCGTTGAGATAGGGGGA	22	58,9	1 (Tm inférieure , Max Poly-X=5)	TGGCCGTGGAGTTTCATAGCTCA	23	58,07	1 (Tm inférieure)
FT_E2E3	421	TCCTGATGCACCCAGTCCAAGTGA	24	59,6	1	GCTTGCTGCAGTTGTTGCTGGA	22	58,79	1 (Tm inférieure)
FT_E3E4	317	TCCAGCAACAACTGCAGCAAGC	22	58,8	1 (Tm inférieure)	GAGCCGCTCTCCCTTTGGCA	20	59,06	1
FT_3UE3	436	TCCAGCAACAACTGCAGCAAGC	22	58,8	1 (Tm inférieure)	CCGGTTTGTAAAGGGCGGGC	20	58,53	1 (Tm inférieure)
KSN_E1E2	317	CGGAACCTTTAGTTGTTGGAAGAGTCA	27	57,2	1 (Tm inférieure)	GGGCCAGGAACATCTGGGTCTG	22	58,89	1 (Tm inférieure)
KSN_E2E3	416	ACCCAGATGTTCCTGGCCCTAGT	23	58,8	Non retenue (Tm inférieure)	TGTGGTGCCTGGAATGTCTGTCA	23	58,19	NON RETENUE Tm inférieure
KSN_E3E4	393	TGTGACAGACATTCCAGGCACCAC	24	59,1	1	CCTGCCTCCTGCTAGCTGCC	20	59,2	1
KSN_FTr1	259	ACCCAGATGTTCCTGGCCCTAGT	23	58,8	1 (Tm inférieure) Couple mélangé avec KSN_FTr1	AGCTCCGGTTTCCTCGTCGT	20	57,81	1 (Tm inférieure)
KSN_FTr2	533	AGGAAGGCTGAAGATTGATTGCCTGG	26	59,1	1 Couple mélangé avec KSN_FTr2	TGTGGTGCCTGGAATGTCTGTCA	23	58,19	1 (Tm inférieure)
NUDX_E1E2	565	GGGAAACGAGACAGTAGTAGTGGCTGA	27	59,5	1 (1 SNP)	AGTGGCTTCGGAAGATTGTCCCAC	24	58,94	1 (Tm inférieure)
NUDX_E1E2b	519	AGTGGCGGTGGTAGTATGCCTGT	23	59,7	1 (1 SNP)	GTGGCTTCGGAAGATTGTCCCACT	24	58,94	1 (Tm inférieure)
PAAS_E1	509	AGAGATCGCCTCTTCCCAACTGACC	25	59,8	1	GGTCCCTGGCAGCCACCATTG	21	59,98	1
PAAS_E1b	521	GGCGGCGGTGTTTTGCATGG	20	60,1	1 (Max Poly-X=4)	ACCCAAAGGCAACAGCAATCCA	22	57,24	1 (Tm inférieure)
PAAS_E1t	581	AGGGCGCAAATTCTTTTAGTTTCAACC	27	57,3	1 (Tm inférieure, Max Poly-X=4)	TGCATGGTCCTGCACCACATTCC	23	59,68	1
BRC1_E1E1	480	GCCGTGCCTTTTCCTCATGACC	22	58,4	1 (Tm inférieure, Max Poly-X=4)	CCGGTCCCTCAGTCCTCGGG	20	60,04	1
BRC1_E1E1b	521	GCAAGAAGGACCGGCACAGCA	21	59,9	1	TGCCTGATCAACATCACCACATGGC	25	59,78	1
BRC1_E1E2	568	AAGATGGGGAGACAGTCAAAGCCA	24	57,8	1 (Tm inférieure, Max Poly-X=4)	CTGTTGTAGACCTCCCATTGTTTTCCA	27	57,15	1 (1 SNP Tm inférieure, Max Poly-X=4)

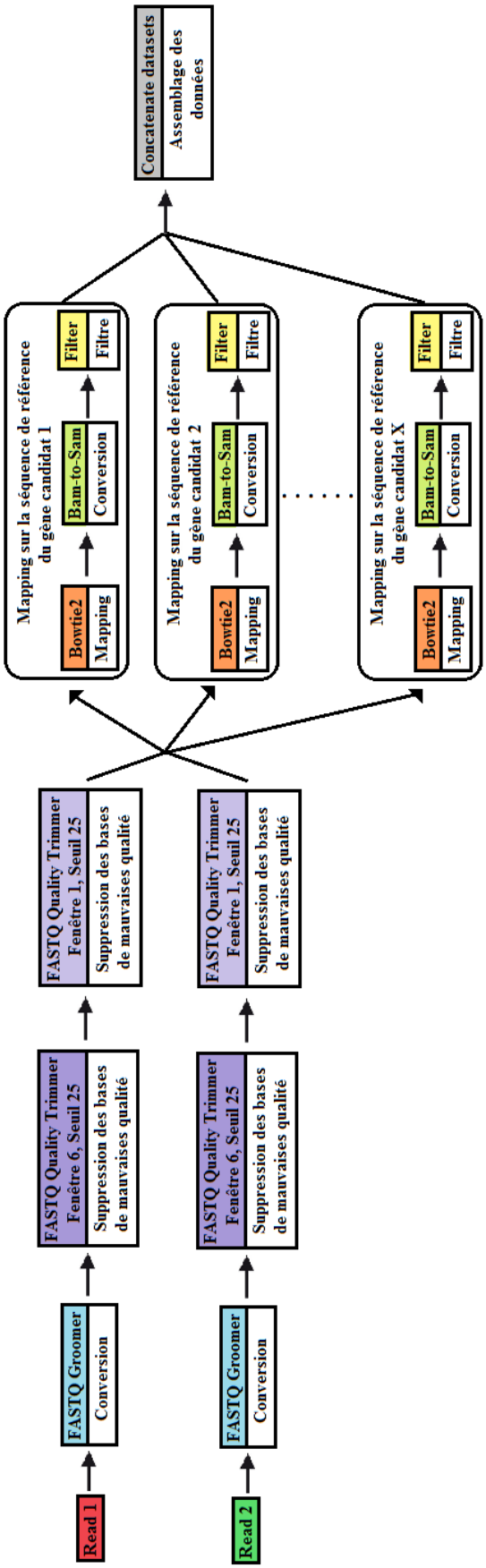
Annexe VIII : Tableau récapitulatif contenant l'ensemble des informations de chacune des amorces mises au point

Nom du couple d'amorce	Taille de l'amplicon	Amorce Forward				Amoce Reverse			
		Séquence	pb	Tm	Information	Séquence	pb	Tm	Information
DFR_5UE1	181	ACTTCCCTCTC TCCCGGAGTC T	22	57,6	3 (Tm inférieure)	AGGGTCTCGC ACGGTGGCTC	20	60,87	3
DFR_E1E2	344	CCGAGTCCGT TTGCGTGACA GG	22	60,3	4	GGCGACATG GAACACTCCG GT	21	59,12	4 (1 SNP en milieu d'amorce)
DFR_E2E3	894	GCGGCGACTC ACTTGACGCT	20	60,1	4	CTTCACTCTCC GGCAAAATTC GACA	25	57,83	4 (1 SNP tout début d'amorce, Tm inférieure, Max Poly-X=4)
DFR_E3E5	493	TGATGTCGAA TTTTGCCGGA GAGTG	25	57,9	3 (Tm inférieure , Max Poly-X=4)	GCGGCCCTCT GCTTTCGGAT	20	59,49	3
DFR_E4E6	680	TGAGCAAGA GGCATGGAA GTTTGCC	25	60,0	2	ACACTGCTCT CATCAACCTC CTGCT	25	59,66	2
DFR_3UE6	264	GCAAAGGGTT TGCTTCCTCCT CC	23	58,1	4 (1 SNP au début (2°) + Tm inférieure)	AGTGAGTCAC GTACCCTCTA ACATTCA	27	57,2	4 (Tm inférieure)
F3'H_E1E1	353	CTCCGGCAAA TCCCAAAGCC ACT	23	59,7	1	TGCCGGAGA ACAGATGGA CGGA	22	59,66	1
F3'H_E1E2	760	CCTTACGGTC CACGGTGGCG	20	59,8	1	ACCCTCCGCC CTACCATCAC C	21	59,7	1
F3'H_E2E2	412	TGCTAGCGCA TGCCTTGGCA AAT	23	60,3	1	GTGTCGGTGA GCTTTCCTCCT TCG	24	59,54	1
F3'H_E2E3	702	CGACATGTTG ACCACTCTGC TCTC	24	57,6	1 (Tm inférieure)	CTCGCGCTAG CATATGAGGG TG	22	57,37	1 (Tm inférieure)
F3'H_E3E3	566	AGCTGAACTC ATTCGGCACC CTCA	24	60,2	1	CCAACCTGGT GCGCGGGT	18	59,88	1
FLS_E1E1	366	GGTCAGAGA ATGAGCAGCC TGGAATCA	27	60,3	2	AGCAGTAGTT AATGGCAGA AGGTGGC	26	59,08	2
FLS_E1I1	479	AAGGGAAGA AAGGGTGGG TGGATCA	25	59,3	4 (EX-IN)	ACTGGAAAAT GCGCGCAAA GGTG	23	59,46	4 (EX-IN, Max Poly-X=4, SNP possible)
FLS_E1I1b	1150	ACTAGAAGG GAAGAAAGG GTGGGTGG	26	59,4	4 (EX-IN)	AGAGTCCGAT GAGACAAAA GGCAACC	26	59,03	4 (Max Poly-X=4, IN-IN SNP possible)
FLS_I1I1	883	CACCTTTGCG CGCATTTTCC AGT	23	59,5	4 (IN-IN SNP possible, Max Poly-X=4)	CAAACCCATG TGATCTCCCC GCA	23	59,19	4 (Max Poly-X=4, IN-IN SNP possible)
FLS_I1I1b	579	AGGCAAGTA GGTGGTTGCC TTTTGT	25	59,0	4 (IN-IN SNP possible, Max Poly-X=4)	GCAAGCAGC CACGTCCACC TT	21	60,18	4 (IN-IN SNP possible)
FLS_I1I1t	599	GAGACCGTTC TAGCTTTGCTC GGG	24	59,4	4 (IN-IN SNP possible)	CACGTCCACT TTGGCTAACA AAGCC	25	59,01	4 (IN-IN SNP possible)
FLS_I1I1q	758	GAAGAGGCTT TGTTAGCCAA CGTGGA	26	59,5	4 (IN-IN SNP)	GGCCCCAGTC GCAGAGACC A	20	60,53	4 (Max Poly-X=4, IN-IN)
FLS_I1I1s	826	ACAGTTCCAC AGCACAAGA GGAGGA	25	59,5	4 (IN-IN)	GCCAAAAAT AGATGCCTCC ACGGCT	25	59,84	4 (Max Poly-X=5, IN-IN)
FLS_I1I1o	801	AGCGTGAGA CCTGATTGGT TACTGGTC	27	60,2	4 (IN-IN)	TGACAATCAG ATCCTCTGCTC ACACCT	27	59,27	4 (IN-IN Plusieurs SNP)
FLS_I1E2	611	TCAAAAGCCA ATCTACACAC ATCCGCA	27	59,3	4 (IN-IN SNP possible, Max Poly-X=4)	CGGCAAGCTT GGAGGCCCTG	20	60,32	4
FLS_E2E3	771	TGTCCCCGCC CTGATCTTGCT	21	59,9	3 (Max Poly-X=4)	ACGAGCTTGG GGTGAGGCCC	20	60,82	3 (Max Poly-X=4)
FLS_5UI1	401	TTTCGGTCTG AGAGAGTGC AAGACA	25	57,9	2 (Tm inférieure)	CCCTCCACGG ACTTAGAGTT CGGA	24	59,23	2
FLS_3UE3	229	GCACAGAAC CACAGTGAGC AAAGAC	25	58,7	2 (Tm inférieure)	GCCTTGTGAG CTCTGAGACA CCC	23	59,2	2

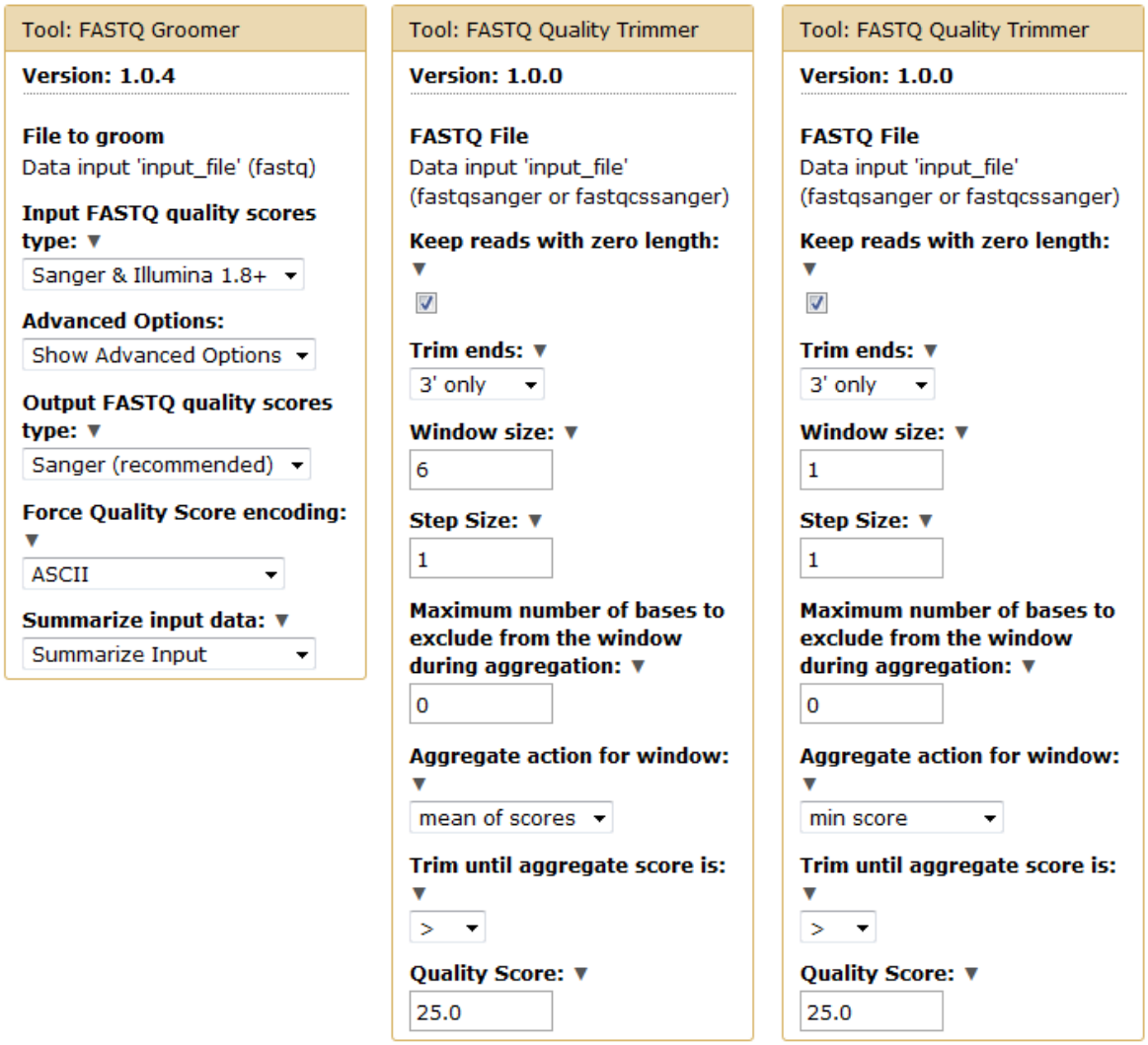
Annexe VIII : Tableau récapitulatif contenant l'ensemble des informations de chacune des amorces mises au point

Nom du couple d'amorce	Taille de l'amplicon	Amorce Forward				Amoce Reverse			
		Séquence	pb	Tm	Information	Séquence	pb	Tm	Information
ZDS_5UE1	493	TGTCTCTCTCG CTTTGTCCTCC TCC	25	59,5	1	GGAGGACCTC TGGGACCTCA CC	22	59,4	1
ZDS_E1E2	707	GTCGATGCCT GGTGGCTCCG	20	59,8	1	ACATGCCTGC AAGTCCAGCT CC	22	59,41	1
ZDS_E2E3	420	AGGGCCGAA GCTGAAAGTG GC	21	59,7	1	AACAACCAA AGAATACGTG GAGTCCCA	27	58,47	1 (Tm inférieure)
ZDS_E3E5	542	AGGCCTTTCA TTGGTGGAAA AGTAGGC	27	59,5	1 (Max Poly-X=4)	AAGCAAGAA TCCCATGTAT GGGGGC	25	59,08	1 (Max Poly-X=5, 1SNP (2°))
ZDS_E3E6	933	AGGCCTTTCA TTGGTGGAAA AGTAGGC	27	59,5	1 (Max Poly-X=4)	CCGCACATCA CTCAATGCTC CATCT	25	59,39	1
ZDS_E6E7	340	TCCTGTCGTC AGGGCTCTGG T	21	59,0	1	CAACGAGCA CTGATGTTGT CACAGTC	26	58,64	1 (Tm inférieure)
ZDS_E7E8	395	CCAAAGGTG GCACACGCAC GA	21	60,4	1	AGGCCTGTGA CATAAGTTTC CCCATCA	27	59,69	1 (1 SNP milieu amorce (8°))
ZDS_E8E10	392	GCTGATGGGG AAACTTATGT CACAGGC	27	59,9	1 (1 SNP fin amorce (22°), Max Poly-X=4)	TCCCTCCACT GGGACGGAA GC	21	59,91	1
ZDS_E10E11	417	CATGCGATGT GCCTGGAATC AAGAGA	26	59,4	1	GCAGGAGTG AACCTTGTCC TTCAATGT	27	59,22	1
ZDS_E11E12	489	GCAATTGAAG CAAGCTTTGG GATTGGA	27	59,1	1	TGGTAGAGGC ATGTAAGGAT CACCTGG	27	59,37	1
ZDS_E12E13	301	TGTGTGTTGA CACCAGGTGA TCCT	24	58,0	1 (Tm inférieure)	ATGGATCTTT ACCAGGTCCT TCACGA	26	57,34	1 (Tm inférieure)
ZDS_E13E14	354	TGGCTTTATTC CCATCATCGC AAGGC	26	59,9	1	TCCCCTCTTGG GATGCAAGTT TCT	24	57,75	1 (Tm inférieure, Max Poly-X=4)
ZDS_3UE13	596	GGCTTTATTCC CATCATCGCA AGGC	25	58,8	1 (Tm inférieure)	GGTAAAAAG AAGCCATGCC CTCATCA	26	57,65	1 (Tm inférieure, Max Poly-X=5)
AG_5UE1	707	CCCTTCTCTGA GTCCCCCTTG C	22	58,3	1 (Tm inférieure, Max Poly-X=5)	GCGTCCAGGA CCGTGTTGGG	20	59,98	1
AG_E1E1	221	GGCCTATGAA AACAAACCC AACACGGT	27	60,1	1 (Max Poly-X=4)	ACTCATAGAG GCGGCCACG GT	21	59,71	1
AG_E1I1	621	TGCTCTGTGA TGCTGAGGTT GCT	23	58,6	1 (Tm inférieure)	AGCAATGGA GGAAGGCAA GGAGC	23	59,19	1
AG_I1I1	559	AGCTCAGCTC CTTGCCTTCCT CC	23	60,2	1	TGGCACTGGG TTTTGATTGGC TGT	24	59,52	1 (Max Poly-X=4)
AG_I1I1b	512	ACAGCCAATC AAAACCCAGT GCCA	24	59,5	1 (Max Poly-X=4)	AGAACTTCCC ATTGGCTCGC AGA	23	58,32	1 (Tm inférieure)
AG_I1I1t	602	ACGAAATCTG CGAGCCAATG GGA	23	59,0	1	ATCTCAATGT GCCAAGGAA GAAACCCA	27	58,73	1 (Tm inférieure)
AG_I1I1q	806	CACAGAACTG GGTTTCTTCCT TGGCA	26	59,4	1	GCAAAGGGTT GGACCTATCT GTGGC	25	59,78	1
AG_I1I1s	486	GCCACAGATA GGTCCAACCC TTTGC	25	59,8	1	AACACAATTG GGGTTGTGTG GCA	23	57,81	1 (Tm inférieure, Max Poly-X=4)
AG_I1E3	650	TGGATGATGT AATGCCACAC AACCC	25	57,6	1 (Tm inférieure)	GCAAAGTGGT TATCTGGGCA CGCA	24	60,23	1
AG_E1E5	823	TGAACGATAC AAGAAGGCA TGTGCAG	26	58,0	1 (Tm inférieure)	TCTGCATGTA CTCAATTTCG GCAAACA	27	57,92	1 (Tm inférieure)
AG_E3E5	388	GCTGCCAAAC TGCGTGCCC	19	59,7	1	CTGCATGTAC TCAATTTCGG CAAACA	26	57,01	1 (Tm inférieure)
AG_E5E6	544	TGTTTGCCGA AATTGAGTAC ATGCAGA	27	57,9	1 (Tm inférieure)	TGCTCGGAGG AGCTGGTTAT TGT	23	57,82	1 (Tm inférieure)
AG_E6E7	314	TTGCACAACA ATAACCAGCT CCTCCG	26	59,7	1	GGGAAATCTG GTCATGGCGC GAG	23	60,18	1
AG_3UE7	407	GCAGGTGGG CATGGAAGCT ACG	22	60,1	1	TGCCAGTACT AACTGTGGGA GAGGTT	26	58,89	1 (Tm inférieure)

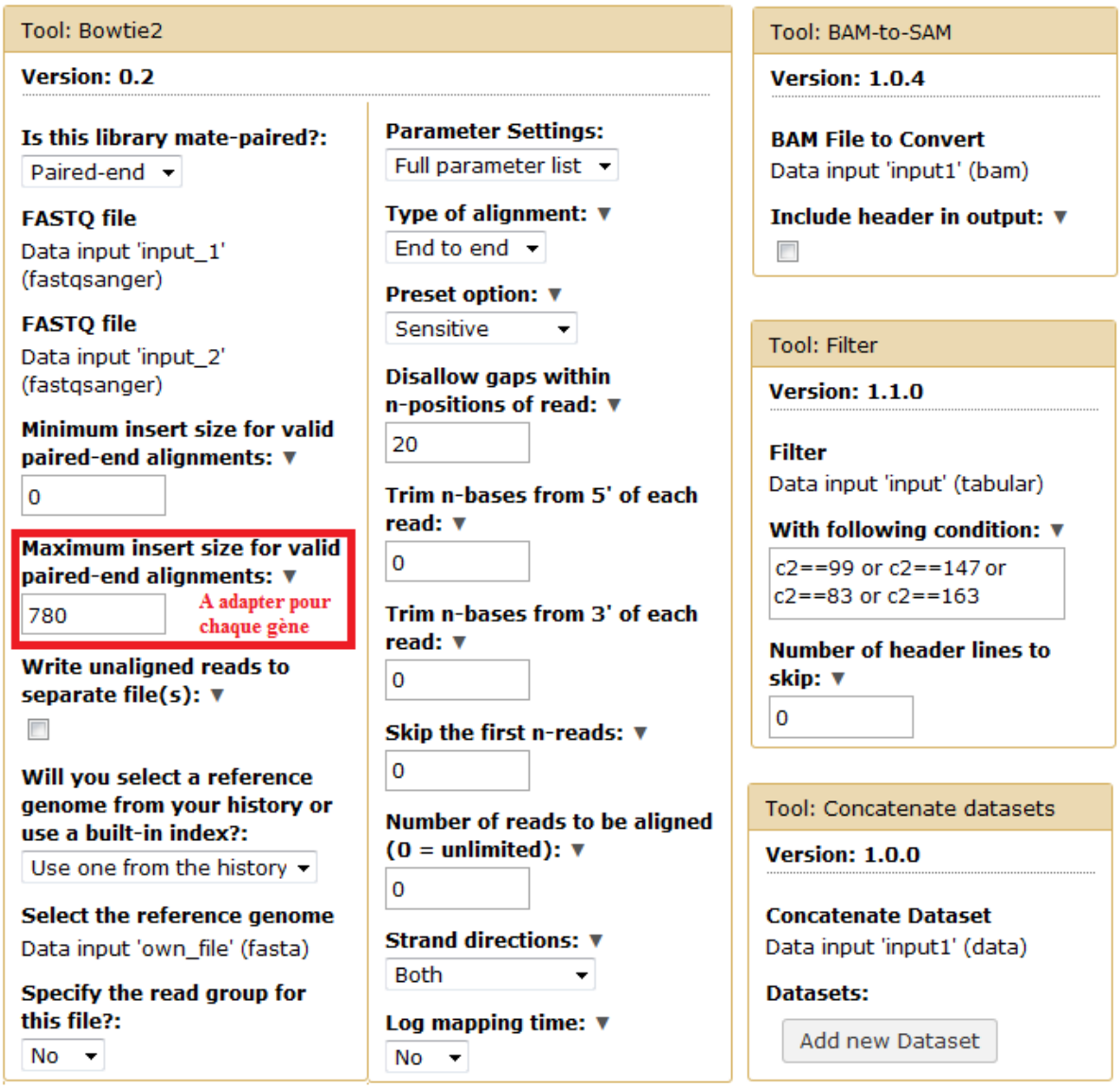
Annexe IX : Schéma du workflow final et paramètres fixés pour chacune des étapes



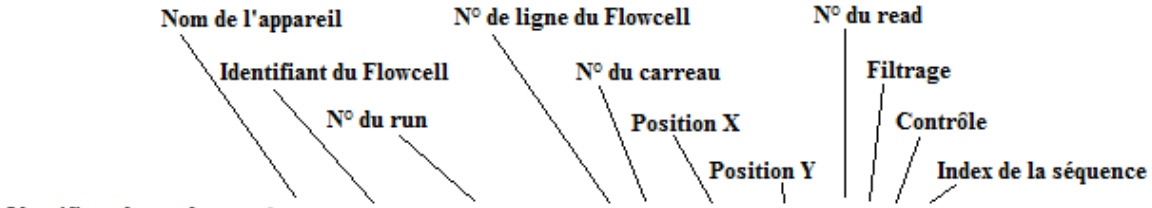
Annexe IX : Schéma du workflow final et paramètres fixés pour chacune des étapes



Annexe IX : Schéma du workflow final et paramètres fixés pour chacune des étapes



Annexe X : Informations présentes dans un fichier de sortie du MiSeq (.fastq)



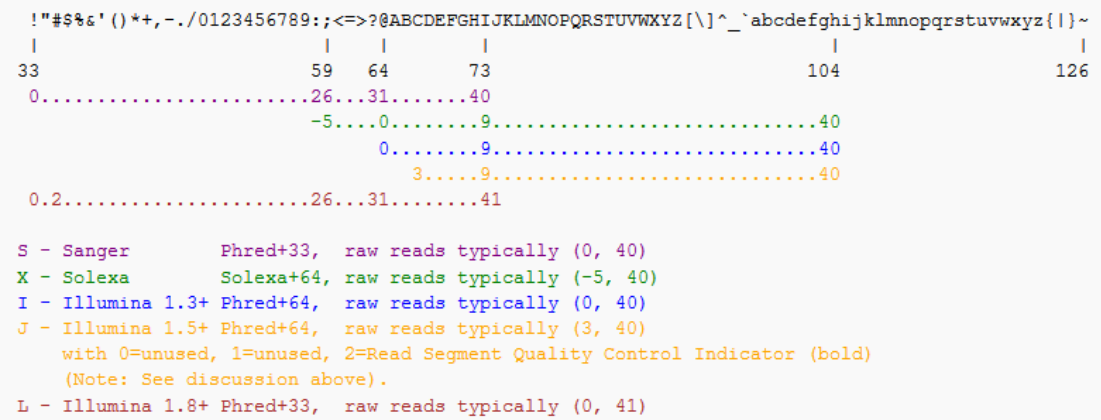
Identifiant du read	@M01075:79:000000000-ACRPA:1:1101:17353:1093 2:N:0:73
Séquence du read	NGTGGCTTCGGAAGCTTGTCCCACTCATACCATCCCCAACCATCACAGAATTCTGGCCC
Identifiant de la qualité	+
Séquence qualité	#-8B@C,6E@6@,+,C;ECCF<FF,C@FFFG@<,CFGCEE7CAF9<F@8FC,CC,6E8F

Remarques :

N°read : 1 pour forward, 2 pour reverse
Filtrage: N indique l'absence de filtrage
Contrôle : 0 indique l'absence de contrôle
Identifiant de la qualité : + indique que l'identifaint du read et de la qualité sont les mêmes

(2015) N. Jacquier

Annexe XI : Les différents modes d'encodage des scores de qualité [33]



Annexe XII : Encodage des bases A, T, G et C et des bases ambiguës [44]

Code	Meaning	Etymology	Complement	Opposite
A	A	**A**denosine	T	B
T/U	T	**T**hymidine/**U**ridine	A	V
G	G	**G**uanine	C	H
C	C	**C**ytidine	G	D
K	G or T	**K**eto	M	M
M	A or C	A**m**ino	K	K
R	A or G	Pu**r**ine	Y	Y
Y	C or T	P**y**rimidine	R	R
S	C or G	**S**trong	**S**	**W**
W	A or T	**W**eak	**W**	**S**
B	C or G or T	not A (**B** comes after A)	V	A
V	A or C or G	not T/U (**V** comes after U)	B	T/U
H	A or C or T	not G (**H** comes after G)	D	G
D	A or G or T	not C (**D** comes after C)	H	C
X/N	G or A or T or C	a**n**y	**N**	**.**
.	not G or A or T or C		**.**	**N**
-	gap of indeterminate length			

Annexe XIII : Tableau complet des flags et de leur signification [48]

FLAG	flags	pair	itself		mate		proper?	aligner?
One of the mate is unmapped								
73	1+8+64 73	1	map	+	unmap			
133	1+4+128 133	2	unmap		map	+		
89	1+8+16+64 89	1	map	+	unmap	-		
121	1+8+16+32+64	1	map	-	unmap	-		
165	1+4+32+128 165	2	unmap	+	map	-		ssaha
181	1+4+16+32+128	2	unmap	-	map	-		bwa
101	1+4+32+64 101	1	unmap	+	map	-		ssaha
117	1+4+16+32+64	1	unmap	-	map	+		bwa
153	1+8+16+128 153	2	map	-	unmap	+		
185	1+8+16+32+128	2	map	-	unmap	-		
69	1+4+64 69	1	unmap	+	map	+		
137	1+8+128 137	2	map	+	unmap	+		
Both unmapped								
77	1+4+8+64 77	1	unmap	+	unmap	+		
141	1+4+8+128 141	2	unmap	+	unmap	+		
mapped in correct orientation and within insert size								
99	1+2+32+64 99	1	map	+	map	-	y	
147	0+1+2+16+128 147	2	map	-	map	+	y	
83	1+2+16+64 83	1	map	-	map	+	y	
163	1+2+32+128 163	2	map	+	map	-	y	
mapped within the insert size but wrong orientation (++ or - -)								
67	1+2+64 67	1	map	+	map	+	y	
131	1+2+128 131	2	map	+	map	+	y	
115	1+2+16+32+64 115	1	map	-	map	-	y	
179	1+2+16+32+128 179	2	map	-	map	-	y	
mapped uniquely but wrong insert size, and could possibly reside in different contigs								
81	1+16+64 81	1	map	-	map	+		
161	1+32+128 161	2	map	+	map	-		
97	1+32+64 97	1	map	+	map	-		
145	1+16+128 145	2	map	-	map	+		
65	1+64 65	1	map	+	map	+		
129	1+128 129	2	map	+	map	+		
113	1+16+32+64 113	1	map	-	map	-		
177	1+16+32+128 177	2	map	-	map	-		

Annexe XIV: Table de conversion des symboles de score de qualité en leur valeur au sein du fichier Excel

0	"	#	$	%	&	'	(	)	[	+	,	-	.	/	{	1	2	3	4	5
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

6	7	8	9	:	;	<	=	>	]	@	A	B	C	D	E	F	G	H	I
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

N. Jacquier

	Diplôme : Ingénieur Spécialité : Horticulture Spécialisation / option : HORVAL Enseignant référent : Agnès Grapin

Auteur(s) : Nathalie Jacquier Date de naissance* : 12/07/1991	Organisme d'accueil : IRHS Adresse : 42 Rue Georges Morel, BP 60057 49071 Beaucouzé Maître de stage : Jérémy Clotault
Nb pages : 50 Annexe(s) : 20	
Année de soutenance : 2015	

Titre français : Sélection de gènes candidats chez des rosiers des XVIIIe et XIX siècles et mise en place d'une stratégie de traitement des données permettant l'étude de leur diversité

Titre anglais : Selection of candidate genes in garden roses of the 18th and 19th centuries and development of a data processing strategy to study their diversity

Résumé (1600 caractères maximum) :

Ce mémoire s'inscrit dans le projet FloRHiGe et plus particulièrement sur l'étude de la diversité de gènes candidats chez des rosiers sélectionnés aux XVIIIe et XIXe siècles, période où la création variétale fut foisonnante. Ce rapport s'est focalisé sur les réflexions et stratégies mises en place en amont de l'étude de la diversité génétique de gènes candidats. Son objectif est d'expliquer le choix de ces derniers et de développer la stratégie de traitement des données mise en place. Dix gènes candidats ont été retenus pour leur rôle dans le déterminisme de caractères d'intérêt. Une stratégie de création d'amorces a alors été élaborée afin d'amplifier de façon spécifique les gènes tout en veillant à respecter des contraintes qui ont été identifiées et hiérarchisées. Le traitement des données sur Galaxy a permis d'obtenir des reads mappés propres et de bonne qualité. La réflexion concernant le choix des paramètres à fixer pour le quality trimming a servi à réaliser un compromis entre la conservation des bases de mauvaise qualité et la perte d'informations due à la suppression de celles-ci. Une méthode de fenêtre glissante a été retenue. Le logiciel de mapping utilisé a été choisi de façon raisonnée en veillant à la bonne gestion des indels et de la position des SNP. Enfin, un code Excel® a été mis en œuvre pour le contigage des reads, l'élimination des erreurs de séquençage et le dosage d'haplotypes.

Abstract (1600 caractères maximum) :

This report is part of the FloRHiGe project and more particularly the diversity study of candidate genes of roses which were selected during the 18th and 19th centuries, period when rose breeding was abundant. This report focused on the reflections and the strategies organized upstream to the diversity study of candidate genes. Its aim is to explain the choice of these genes and to develop the strategy of data processing organized. Ten candidate genes were selected for their role in the determinism of interest characters. Then a strategy of primers creation was developed to amplify in a specific way the genes while respecting constraints which were identified and ranked. The data processing on Galaxy allowed us to obtain clean mapped reads of good quality. The reflection concerning the choice of the parameters to be fixed for the quality trimming served to realize a compromise between the preservation of the poor quality bases and the loss of information due to the deletion of these. A method of sliding window was retained. The mapping software was chosen in a reasoned way by watching the good management of indels and SNP's positions. Finally, an Excel® code was created for the reads assembly, the elimination of the sequencing errors and the haplotypes countig.

Mots-clés : Roses, gènes, sélection, stratégie, amorces, mapping, haplotype, quality trimming.

Key Words: Roses, genes, selection, strategy, primers, mapping, haplotype, quality trimming.

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*