MONTPELLIER SUPAGRO
Centre International d'Etudes Supérieures en Sciences Agronomiques

THESE

pour obtenir le grade de

DOCTEUR EN BIOLOGIE

ECOLE DOCTORALE SIBAGHE

Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement

BASES MOLECULAIRES DE LA VARIATION CLONALE CHEZ LA VIGNE (*Vitis vinifera* L.)

APPROCHE PANGENOMIQUE

par

Grégory CARRIER

Présentée le 13 décembre 2011 devant le jury composé de

M. H.Quesneville	Directeur de Recherche, INRA, Versailles	Rapporteur
M. P.Sourdille	Directeur de Recherche, INRA, Clermont-Ferrand	Rapporteur
Mme C. Vitte	Chargée de Recherche, CNRS, Gif-sur-Yvette	Examinatrice
M. O. Viret	Chef de Département, Station ACW, Suisse	Examinateur
M. J.C. Glaszmann	Directeur de Recherche, CIRAD, Montpellier	Président
M. P. This	Directeur de Recherche, INRA, Montpellier	Directeur de thèse
M. J.M. Boursiquot	Maitre de Conférences, Montpellier SupAgro	Invité
M. L. Le Cunff	Ingénieur IFV, Le Grau-du-Roi	Invité

Je dédie ces travaux à mon Grand père

Resumé

L'exploitation de la variation clonale est une des voies d'amélioration utilisée chez un grand nombre de plantes d'intérêts agronomiques telles que la pomme de terre, le café et la vigne. En effet, après plusieurs cycles de reproduction végétative, des caractéristiques agronomiques stables apparaissent donnant naissance à une diversité phénotypique remarquable appelée « diversité clonale ». Chez la vigne, cette diversité clonale est d'une importance majeure pour les viticulteurs puisqu'elle permet une amélioration variétale sans changer d'identité de cépage en conformité avec la réglementation fixée par les Appellations d'Origine Protégée.

L'hypothèse la plus parcimonieuse expliquant cette diversité phénotypique clonale est l'accumulation de mutations somatiques au cours des cycles de reproduction végétative. L'objectif de cette thèse a été de dresser un panorama le plus exhaustif possible des différents polymorphismes moléculaires entre les génomes de plusieurs clones. Dans un premier temps, trois clones de Pinot ont été séquencés par la technique 454 GS-FLX puis, dans un second temps 11 clones de quatre cépages ont été séquencés avec la technique Illumina HiSeq 2000. Afin d'analyser la grande quantité de données obtenues, nous avons construit un pipeline d'analyse (Bacchus pipeline) permettant d'identifier tous les types de polymorphismes moléculaires entre les différents génomes.

Nos résultats permettent pour la première fois de réaliser un inventaire exhaustif des polymorphismes moléculaires dans un contexte de multiplication végétative. L'ensemble des mutations polymorphes entre deux clones a pu être identifié : SNPs, indels (2,5 SNPs et 11,5 indels par Mb en moyenne) ainsi que des variations d'ordre structural (larges insertions ou délétions) représentant la classe de mutations la plus fréquente (129 évènements par Mb entre deux clones en moyenne). Afin d'évaluer le polymorphisme d'insertion généré par ces éléments nous en avons étudié quatre par une approche S-SAP sur plusieurs niveaux de diversité (inter-espèces, inter-cépages, inter-clones et entre plusieurs tissus d'un même individu). L'analyse phylogénétique au niveau des espèces est conforme à celle réalisée avec d'autres types de marqueurs moléculaires (SSR, SNP). Cependant, une forte instabilité de ces insertions a été confirmée entre les clones et entre les tissus d'un même d'individu.

L'identification des clones par une méthodologie moléculaire serait d'une grande importance pour la filère. Nos résultats indiquent que les mutations les plus pertinentes pour la mise en place d'une méthodologie d'identification des clones sont les mutations de types SNP et petits indels qui sont certes moins fréquentes que les variations structurales mais plus stables.

Remerciements

Je tiens à remercier tous les membres de mon jury d'avoir accepté d'évaluer mon travail.

Merci à l'ANRT, à Jean-Pierre Van Ruyskensvelde et aux partenaires de la sélection vigne pour m'avoir fait confiance et financé ma thèse.

Merci du fond du cœur à Patrice This, qui m'a accueilli depuis 4 ans dans son équipe et de m'avoir fait confiance tout au long de mon stage de Master et de ma thèse. Merci pour ta gentillesse et pour tous tes conseils professionnels mais aussi humains qui m'ont permis d'avancer et de m'épanouir dans ton équipe.

Merci à Jean-Michel Boursiquot, pour tous tes conseils, ta bonne humeur, et la confiance que tu m'as apporté. Merci de m'avoir fait découvrir l'ampélographie, je n'aurai jamais cru qu'autant de nuances de vert existaient !

Un très grand merci à Loic Le Cunff, encadrant et ami, qui a su me faire confiance et qui m'a encouragé et conseillé tout le long de cette belle aventure qu'est la thèse. Celle-ci aurait bien été différente sans toi.

Un très grand merci à Laurent Audeguin pour ta bonne humeur et à ta façon très humaine d'animer l'UMT Génovigne. Merci à Isabelle et Sylvie pour m'avoir aidé dans tous les déboires administratifs. Merci à toute l'équipe de l'IFV.

Merci aux partenaires de la sélection vigne pour leur accueil chaleureux, leurs conseils et leurs aides. Merci de m'avoir fait goûter d'aussi bons vins et de m'avoir transmis tant de photos de vos clones.

Merci à Anne Françoise Adam Blondon, François Sabot, Franc Christophe Baurens, Romain Guyot, de m'avoir suivi pendant ces trois ans. Vos conseils ont été d'une grande aide pour faire les bons choix.

Merci à toute l'équipe Viti, merci à Thierry pour toutes les photos et tous tes conseils et encouragements. Merci à Roberto de m'avoir fait confiance pour encadrer tes petits stagiaires, merci à Amandine pour ta bonne humeur, tes confitures et tes extractions d'ADN ! Merci à Charles pour tes conseils et ton aide pour la fonction des gènes. Merci à Jean Pierre pour ton aide bibliographique et les débats écolo. Merci à Philipe pout toutes ces corrections d'anglais surtout dans le dernier rush de la thèse !

Merci à Delphine pour ta bonne humeur, tes cookies et ton fameux jeu de quilles Finlandais !

Un grand merci au bocal thésard ! Yung Fen pour m'avoir supporté pendant trois ans et toutes les discussions super enrichissantes que nous avons eu, Agota pour ta bonne humeur et ton fameux « Unicum » ! Merci à Pilar pour toutes ces petites attentions et la bonne humeur que tu transmets dans le bureau. Merci à Catherine pour tous tes conseils. Merci à Alexis de m'avoir montré et aidé en perl, sans toi ça aurait été beaucoup plus dur ! Un merci aux anciens précaires, Stéphane et Alex qui m'ont beaucoup apporté.

Merci à Pierre Bourget et à Vincent Maillol pour avoir été des supers stagiaires et m'avoir beaucoup aidé pendant ma thèse.

Merci à toutes les girls de l'AMM pour leur bonne humeur et leurs encouragements, ainsi qu'à Sylvain pour tous tes conseils et ton aide tout le long de la thèse.

Merci à mes p'tits ordis de ne m'avoir jamais lâché ! Merci à Bertrand, à Jean François et à Manuel de nous avoir accueilli sur votre serveur et d'avoir toujours été là quand ça plantait. Désolé pour le remplissage des disques… Merci à mon logiciel de retouche d'images préféré : Paint.

Merci à l'équipe de la GénoToul de Toulouse pour leur accueil chaleureux, leur expertise et leur bonne humeur.

Merci à Marguerite de m'avoir si bien accueilli et pour tes superbes photos de noyaux.

Merci à Marc André Selosse pour m'avoir fait confiance pour donner des cours à l'UMII.

Merci à Josette pour m'avoir donné le gout de la science il y a 12 ans et de m'avoir accompagné jusque-là.

Un grand merci à mes parents, Lyne et Richard, pour tous leurs encouragements, leur soutien et leur amour.

Un grand merci à ma p'tite Maud pour m'avoir accompagné et soutenu tout le long de ma thèse. Merci pour toutes tes corrections orthographiques, merci de me redonner confiance quand j'en ai besoin.

Merci au Crang, P'tit frère, Grand pas, JC, Pol et aux deux artistes Ptit moine et Picsou… Merci à Mathieu pour toutes ces belles photos, Merci aux amis et à la famille pour vos encouragements de tous les jours.

Merci à toutes les personnes que j'ai pu rencontrer et qui m'ont accordés de leur temps au cours de cette aventure.

Glossaire

454 : Séquenceur nouvelle génération développé par la société Roche
ADN : Acide DésoxyriboNucléique
AFLP : Amplified Fragment Length Polymorphism
ANRT : Association Nationale de la Recherche et de laTechnologie
AOC : Appellation d'Origine Contrôlée
AOP : Appellation d'Origine Protégée
ARN : Acide RiboNucléique
BAC : Bacterial artificial chromosome
BAM : Fichier SAM compressé
CNV : Copy Number Variation
CTPS : Comité Technique Permanent de la Sélection des plantes cultivées
EMS : EthylMéthyl Sulfonate
FAO : Food and Agriculture Organization
FASTA : Format de fichier contenant les informations de séquences
FASTQ : Format de fichier contenant à la fois les informations des fichiers Fasta et Qual
Gb : Gigabases (10^9 bases)
GFF : Format de fichier contenant des informations d'annotation ou de polymorphisme
IFV : Institut Français de la Vigne et du Vin
IGP : Indication Géographique Protégée
Illumina HiSeq 2000 : Séquenceur nouvelle génération développé par la société Illumina
INAO : Institut National de l'Origine et de la Qualité
Indel : Insertion ou délétion
INRA : Institut National de la Recherche Agronomique
Kb : Kilobases (10^3 bases)
LTR : Long Terminal Repeat
M : Mole
Mb : Mégabases (10^6 bases)
NGS : New generation sequencer
OIV : Organisation Internationale de la Vigne et du Vin
Pb : paire de bases
PCR : Polymerase Chain Reaction
QUAL : Format de fichier contenant les informations de la qualité des séquences
SAM : Format de fichier contenant les informations de séquences alignés
SNP : Single Nucleotide Polymorphism
S-SAP : Sequence-Specific Amplification Polymorphism
SSR : Single Sequence Repeat
TE : Transposable Element
VATE : Valeur Agronomique, Technologique et Environnementale

Sommaire

Avant-propos

La vigne (*Vitis vinifera* L.) est l'une des plus importantes espèces fruitières cultivées dans le monde. Avec une production totale de raisins estimée, en 2009, à 67 millions de tonnes pour une valeur économique de 32 milliards de dollars (statistiques F.A.O), la vigne est la 14ème culture d'importance économique au niveau mondial alors qu'elle ne représente seulement que 0,5% des terres cultivées. La majorité de la production de la vigne (71%) est transformée en vin, produit à haute valeur ajoutée possédant une symbolique gastronomique religieuse et sociale qui n'a pas son pareil. La France se situe à la deuxième place au niveau mondial en terme de production vinicole après l'Italie avec une production de 41 millions d'hectolitres. Les vins français sont la référence internationale de par leur typicité, leur qualité et leur histoire.

La vigne *Vitis vinifera* L. est cultivée depuis l'Antiquité et comprend plus de 6000 variétés, dont les plus anciennes seraient âgées de centaines, voire de milliers d'années. La multiplication végétative, qui fait partie du processus de domestication de nombreuses espèces, a permis très tôt de multiplier des cépages sélectionnés pour leurs caractères d'intérêts agronomiques. Tout en conservant la typicité et l'identité du cépage, les individus multipliés ont acquis au fil du temps une certaine originalité phénotypique donnant naissance à une diversité clonale remarquable. La diversité clonale est d'une importance majeure pour les viticulteurs. La réputation des cépages ainsi que la réglementation fixée par les A.O.P. (54% des domaines) contraignent les viticulteurs à conserver l'identité du cépage implanté dans leur domaine. La diversité clonale est alors généralement le seul moyen leur permettant d'accéder à de la diversité végétale.

L'objectif de cette thèse est de comprendre l'origine de la diversité phénotypique clonale. L'hypothèse la plus vraisemblable expliquant cette diversité phénotypique clonale est l'accumulation au cours du temps de mutations somatiques. Nous avons dressé un panorama le plus exhaustif possible des différents polymorphismes moléculaires impliqués dans cette diversité. L'étude de la fréquence et de la stabilité de ces mutations somatiques permettra de

définir des stratégies en vue de la mise en place d'une méthodologie d'identification et de sélection des clones.

Ma thèse a débuté en décembre 2009 à l'UMT Génovigne®, unité qui regroupe une partie du pôle végétal de l'IFV (Institut Français de la Vigne et du Vin) et une partie de l'équipe DAVEM (Diversité et Adaptation de la Vigne et des Espèces Méditerranéennes) de l'INRA (UMR AGAP) et de Montpellier SupAgro. Elle a été financée par une bourse CIFRE (IFV, les partenaires de la sélection clonale) et par l'ANRT (Association Nationale de Recherche et Technologie). Elle a été co-encadrée par Patrice This, Jean-Michel Boursiquot et Loïc Le Cunff.

Chapitre 1,

Introduction

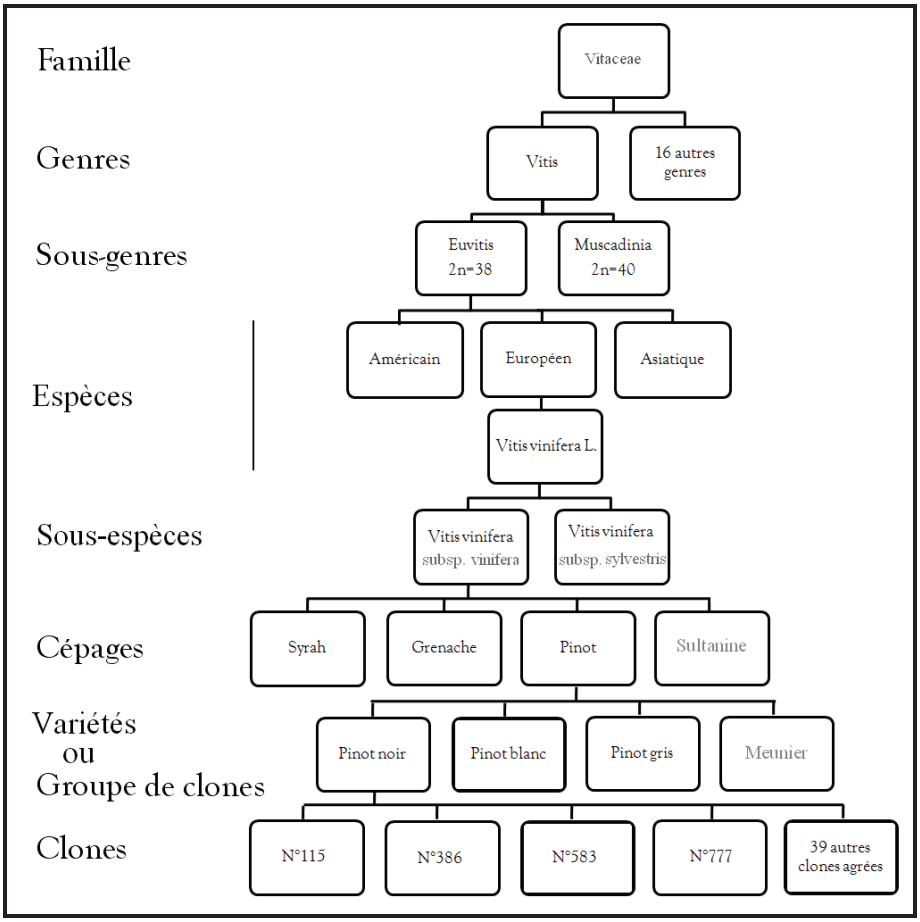


Figure 1 : Taxonomie de la vigne.

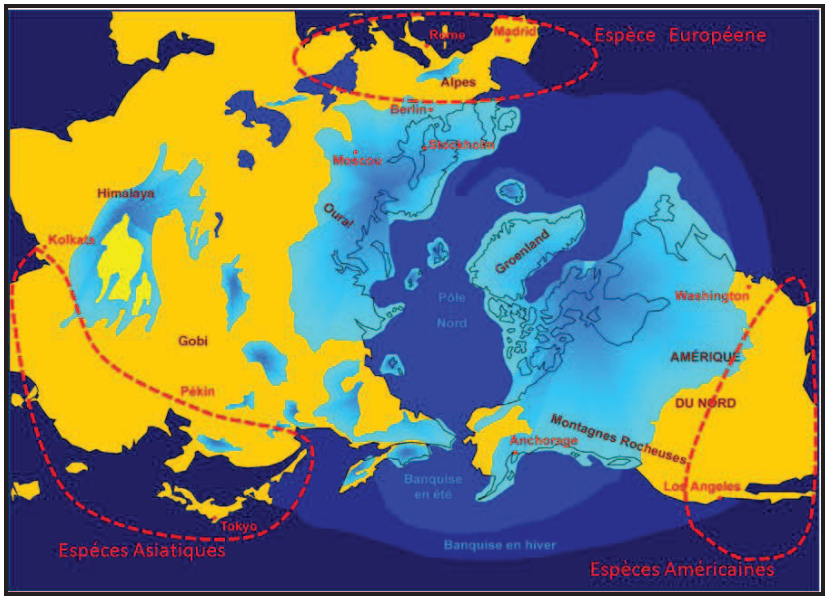


Figure 2 : Hypothèses de spéciation des espèces du genre *Vitis* durant la dernière glaciation (d'après Peros *et al.*, 2010).

1-La vigne

La vigne traditionnellement cultivée appartient à l'espèce *Vitis vinifera* L. C'est une espèce à vocation agronomique remarquable, chargée d'histoire et de symboles, et offrant une diversité exceptionnelle (This *et al.*, 2006).

1.1-Taxonomie et utilisation de la vigne

La vigne appartient à la famille des *Vitaceae* (The Angiosperm Phylogeny Group, 2009) qui comporte 17 genres (Figure 1), parmi lesquels le genre *Vitis* qui signifie « baguette courbée » en grec ancien (Gaffiot, 1934). D'après Péros *et al.* (2010), l'origine du genre *Vitis* est située en Eurasie et il s'est ensuite étendu vers l'Ouest sur le continent américain. La séparation des continents et les périodes de glaciations successives du pléistocène, ont alors provoqué l'isolement de populations qui a conduit à des événements de spéciations (Figure 2 ; Péros *et al.,* 2010). Le genre *Vitis* est en fait composé de deux sous-genres (*Euvitis et Musacadinia*) sur la base des caractères morphologiques et anatomiques (Galet, 1993) ainsi que de leur garniture chromosomique : 2n=40 pour *Muscadinia* et 2n=38 pour *Euvitis* (Bouquet, 1982). Le sous-genre *Muscadinia* comprend deux espèces (*Muscadinia popenoei,* et *Muscadinia rotundifolia*) originaires du Mexique et du Sud-Est des Etats-Unis. Seule l'espèce *Muscadinia rotundifolia* est cultivée. Cette espèce est d'un grand intérêt pour les programmes d'amélioration végétale. En effet, elle présente des niveaux de résistances élevées à un très grand nombre de maladies telles que le mildiou et l'oïdium (Bouquet, 1982). Le sous-genre *Euvitis* comprend une soixantaine d'espèces diploïdes dont *Vitis vinifera* L. Ces espèces, malgré la spéciation, sont restées inter-fertiles. Elles sont classées selon leurs origines géographiques (Figure 1 ; Galet, 1988)) : 1) les vignes américaines, utilisées en particulier depuis la crise phylloxérique comme porte-greffes ou en croisement avec *Vitis vinifera* L. (Levadoux, 1956) ; 2) les vignes asiatiques notamment l'espèce *Vitis amurensis* qui est utilisée dans les programmes d'amélioration pour sa tolérance au froid (Olmo, 1976) ; 3) la vigne euro-asiatique, *Vitis vinifera* L. qui regroupe l'ensemble des cépages cultivés, de cuve et de table appartenant à la sous espèce *Vitis vinifera* subsp. *vinifera,* ainsi que les vignes sauvages de la sous espèce: *Vitis vinifera* subsp. *sylvestris*.

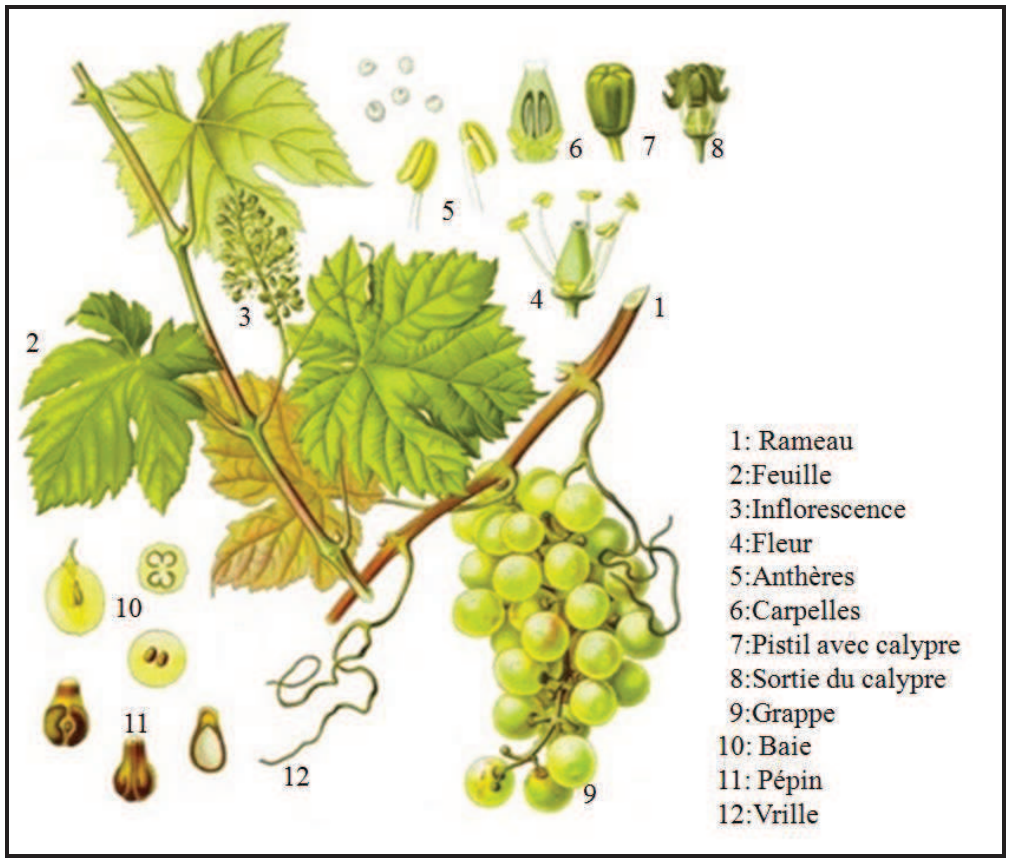


Figure 3 : La vigne, présentation des différents organes.

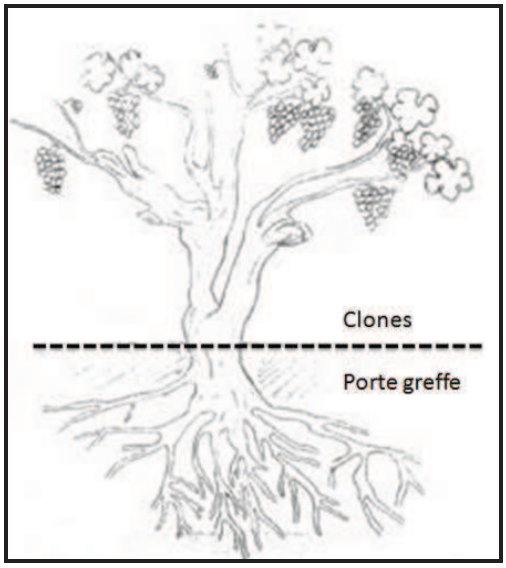


Figure 4 : Cep de vigne cultivée présentant l'assemblage porte-greffe / greffon.

1.2-Description botanique

Vitis vinifera L. est une plante pérenne à port grimpant. Les feuilles ont une distribution alterne et sont pétiolées avec cinq nervures principales. Elles possèdent un sinus pétiolaire et sont plus ou moins découpées, constituées d'un ou plusieurs lobes. Les tiges lignifiées sont appelées sarments ; ils peuvent atteindre une très grande longueur et sont capables de se fixer grâce aux vrilles dont ils sont pourvus. Les fleurs sont à 5 pétales, petites, de couleur verte, formant des inflorescences en grappe. Les fruits sont charnus et communément appelés baies de raisin (Figure 3 ; Viala & Vermorel, 1910; Galet, 1993).

Depuis la crise phylloxérique dans les années 1870, le recours au greffage s'est quasiment généralisé et un plant de vigne est désormais composé d'un porte-greffe, obtenu à partir d'espèces d'origines américaines, pour la partie racinaire et d'un cépage de *Vitis vinifera L.* pour la partie aérienne (Figure 4 ; Legros, 1997).

1.3-Reproduction

La vigne peut se reproduire par voie sexuée ou par voie végétative (Viala & Vermorel, 1910). Pour maintenir les caractères du cépage, la vigne cultivée est multipliée essentiellement par voie végétative ou asexuée. La reproduction sexuée est utilisée pour la création de nouveaux cépages.

1.3.1-Sexuée

La vigne cultivée est majoritairement hermaphrodite, à cycle reproductif long. Il s'écoule en général entre 3 et 5 années pour qu'un nouvel individu produise de nouveaux pépins. Le mode de reproduction de la vigne n'est pas toujours bien déterminé. On suppose que la pollinisation est principalement anémophile (Galet, 1993). La vigne cultivée est vraisemblablement à la fois allogame et autogame, bien que les individus issus d'autofécondation soient en général peu viables (Levadoux *et al*., 1956). En effet, la vigne présente une très forte dépression de consanguinité et supporte généralement très mal l'autofécondation (Valleau, 1916). La reproduction sexuée a été à l'origine de la diversification variétale (Boursiquot & This, 1996), et a permis de générer de nouveaux cépages. Ainsi par exemple, les croisements entre le Pinot et le Gouais ont donné naissance à plus de 20 cépages, dont le Chardonnay ou encore le Gamay (Bowers *et al*., 1999).

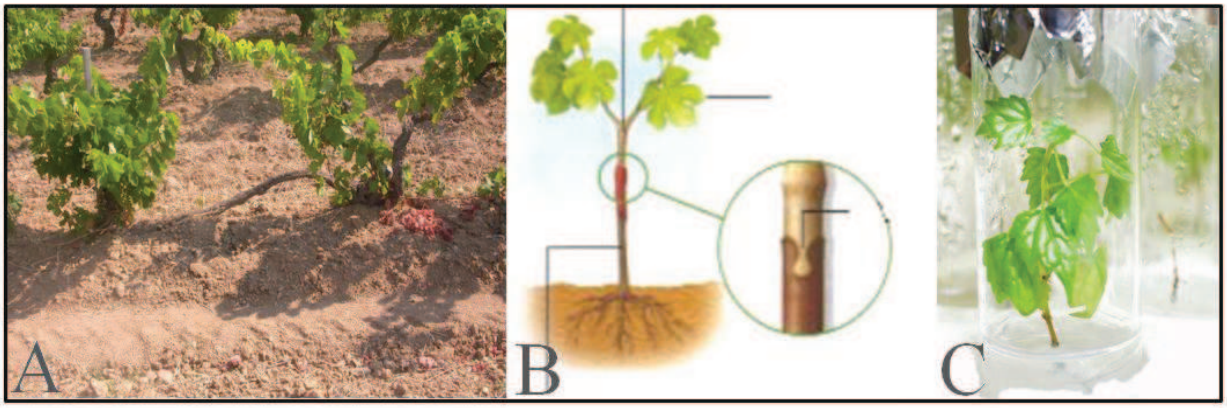


Figure 5 : Différentes méthodes de multiplication végétative pour la vigne. A) Marcottage, B) Greffage, C) Culture *in vitro*.

1.3.2-Asexuée

A l'état sauvage, la vigne peut se multiplier par voie végétative sous la forme de marcottes (Figure 5). Une partie du sarment enterré va être capable de se bouturer et de régénérer un nouveau système racinaire (Levadoux, 1956). En viticulture, la reproduction végétative est très utilisée car elle permet la multiplication et la conservation des différents cépages sélectionnés. Elle permet également une homogénéité de culture et le maintien de la typicité du cépage. Certains cépages très anciens possédant des qualités particulières ont ainsi pu être conservés. C'est le cas par exemple du Muscat à petits grains, de la Sultanine ou du Pinot.

Après la crise phylloxérique, les procédés de multiplication comme le bouturage et le marcottage ont été abandonnés pour être remplacés presque exclusivement par le greffage (Figure 5 ; Pouget, 1990). Celui-ci peut être réalisé de plusieurs manières : en fente, à l'anglaise ou en oméga. De nouveaux procédés de multiplication par culture *in vitro* ont été mis au point chez la vigne il y a une trentaine d'années (Bouquet *et al.*, 1989). Ces techniques sont peu utilisées par les pépiniéristes pour la multiplication dans la pratique. Cependant, c'est un moyen efficace pour restaurer du matériel sain à partir de plantes malades que l'on souhaite sauvegarder.

1.4-Le génome de la vigne

Le génome de *Vitis vinifera* L. est composé de 19 paires de chromosomes (n=38). La taille du génome est évaluée à 470 Mb (Séquençage PN40024, 12X), similaire à celle du peuplier (485 Mb) et du riz (430 Mb) mais environ quatre fois plus importante qu'*Arabodipsis thaliana* (125 Mb). La taille des chromosomes varie de 16,6 Mb pour le chromosome 17, à 29,7 Mb pour le chromosome 14. Actuellement, deux individus de *Vitis vinifera* L. ont été séquencés avec une stratégie Whole Shotgun (Edwards & Caskey, 1991).

1) L'individu PN40024 par le consortium international Franco-Italien (Jaillon *et al.*, 2007), considéré comme la séquence de référence de la vigne, actuellement en version 12X (12-Fev-2010 http://www.genoscope.cns.fr/).

2) Le clone Pinot noir ENTAV-INRA® n°115, séquencé par Velasco *et al.* (2007) en 6,4X.

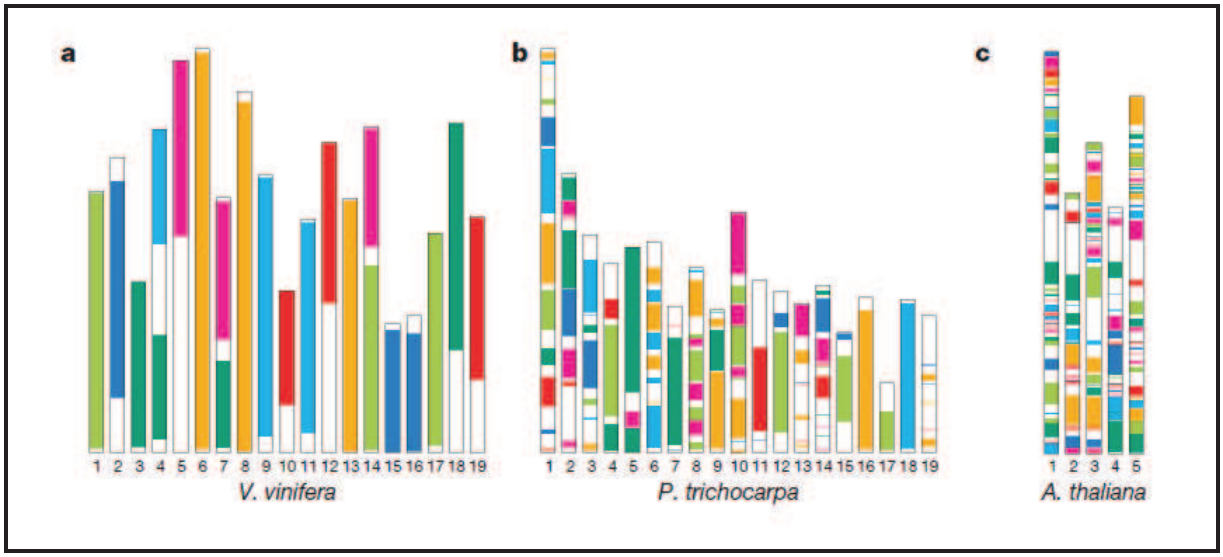


Figure 6 : Comparaison des régions paralogues entre 3 espèces modèles : la vigne (*V. vinifera),*le peuplier (*P. trichicarpa)* et l'arabette *(A. thaliana*). (Jaillon *et al*., 2007).

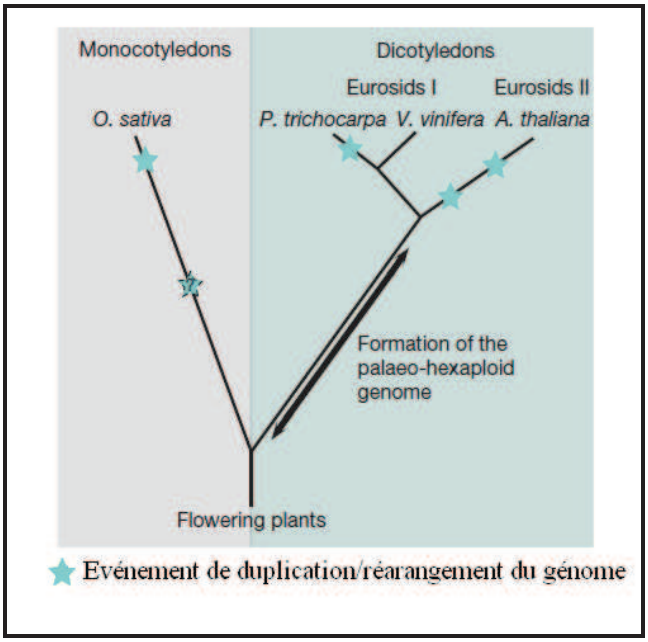


Figure 7 : Evénements de duplication et de réarrangement du génome chez les espèces végétales (Jaillon *et al*., 2007).

1.4.1-Le génome de référence de la vigne

L'individu PN40024 est issu de neuf générations d'autofécondation d'un Pinot noir. Cependant, lors du 4éme croisement, du pollen issu du cépage Helfensteiner (issu d'un croisement entre le Pinot et le Frankenthal) a vraisemblablement été utilisé (Jaillon et *al.* 2007). L'homozygotie de cet individu est évaluée à 97% et il a été séquencé essentiellement par la technique Sanger (Edwards & Caskey, 1991). Les séquences obtenues ont été assemblées en contigs puis en super contigs (scaffolds) et ordonnées en s'appuyant sur les cartes génétiques (Adam-Blondon *et al.*, 2004; Doligez *et al.*, 2006). Dans sa dernière version (12-Fev-2010) les scaffolds positionnés représentent 406 Mb. Les scaffolds non orientés (dit random), où un seul marqueur moléculaire a pu être identifié sur les cartes génétiques, représentent 31 Mb. Le chromosome unknown, regroupe les scaffolds non situés dans le génome, et représente 23 Mb. Les régions de connexions entre les scaffolds ont été indiquées par un ajout empirique de nucléotide N. Ces régions représentent 10 Mb. Du fait de la présence d'un grand nombre de séquences répétées dans les régions télomériques et centromériques, ces régions ont été difficilement assemblées.

Le génome est composé de 32,8% de GC, 0,82% de CpG et 2,19 % de CnG. L'annotation de la séquence a permis de prédire le nombre de gènes contenus dans le génome de la vigne, soit 23 346 gènes prédits (19 mars 2010). On estime que le génome contient 6,3% de régions codantes et 41% d'éléments répétés. L'annotation des éléments répétés a été faite à l'aide de différents logiciels selon le type d'élément (Sputnik pour les SSRs, ReAS et RepeatMasker pour les éléments transposables (Jaillon *et al.,* 2007)). On estime que le génome de la vigne contient en moyenne 480 séquences microsatellites (SSRs) par Mb et est constitué d'au moins 17% d'éléments transposables (Jaillon *et al.,* 2007).

Le génome de la vigne comporte un grand nombre de régions paralogues (Figure 6), résultat d'événements de polyploïdisation. Ces résultats ont permis de déterminer que le génome de la vigne provient d'un ancien polyploïde constitué de trois génomes ancestraux (hexaploïde). Ce génome ancestral est également commun à l'arabette (*A. Thaliana*) et au peuplier (*P. trichocarpa*), et à la différence de la vigne, ceux-ci ont subi par la suite d'autres événements de duplication/réarrangement (Figure 7).

Figure 8 : Fresque sur la viticulture Egyptienne (tombe de Mererou-Ka, 2345 av. JC).

Figure 9 : Représentations de Dionysos et Bacchus au cours du temps.

1.4.2-Le génome du Pinot noir ENTAV-INRA® n°115

Les séquences ont été obtenues par séquençage de banques BAC en utilisant la méthode de Sanger et la technologie 454-GS20. L'assemblage a permis de former 58 611 contigs dont 81% ont pu être positionnés sur une carte génétique (Troggio *et al.*, 2007). Le séquençage de cet individu a permis d'avoir une idée de l'hétérozygotie de la vigne. Celle-ci est très importante, deux millions de SNPs environ ont été identifiés, ainsi qu'un million de petits indels. Ainsi, environ 1 SNP toutes les 250 bases différencient les deux haplotypes du Pinot noir n°115 (Velasco *et al.,* 2007).

1.5-Histoire de la domestication de la vigne

Bien avant sa domestication, la vigne sauvage était exploitée par l'Homme. Celui-ci cueillait et consommait directement le raisin (Mc Govern, 2003). La domestication de la vigne serait apparue au cours du Néolithique (5000 à 6000 ans av. JC.) lorsque l'Homme se sédentarisait et développait l'agriculture dans les régions du Sud du Caucase et du croissant fertile (Mc Govern, 2003). A partir de ces régions, la viticulture va s'étendre au fil des civilisations autour du bassin méditerranéen et au Moyen-Orient. Vers l'Est, en Asie, ce sont les Perses et les Phéniciens qui propageront la vigne. Vers l'Ouest, la viticulture s'implante en Egypte (Figure 8) puis en Grèce et en Italie où les Romains la diffuseront à l'ensemble de la Méditerranée. En Gaule, les premiers vignobles sont implantés par les phocéens à Massalia (Marseille, 600 av. JC, (Dion, 1982)). Les Romains vont ensuite implanter la viticulture dans le Languedoc et le couloir Rhodanien (125 av. JC.) avec le cépage *Allobrogica*, éventuel ancêtre du cépage Syrah ou encore le *Biturca* celte, hypothétique ancêtre du Cabernet franc (Levadoux, 1956). Ces données indiquent que les Romains pratiquaient déjà une sélection des meilleurs cépages qu'ils multipliaient de façon végétative tout en maintenant dans certains cas une multiplication sexuée. Portée par le christianisme, la viticulture va ensuite petit à petit se répandre dans le monde entier. Les vignobles du Nouveau Monde (Amérique, Australie, Afrique du Sud, Nouvelle Zélande) ont été établis dans un premier temps par les missionnaires qui ont apporté des pépins issus de croisements (Boursiquot & This, 2000 ; This *et al*., 2006) puis par les colons chrétiens durant les XVII et XVIIIème siècles avec des boutures (Royer, 1988). Aujourd'hui la vigne est cultivée sur les 5 continents.

Figure 10 : Le phylloxera sous ses différentes formes. (Galles phylloxériques, Photographie de droite).

Au cours de la domestication, les Hommes ont sélectionné des caractères agronomiques leur permettant d'accroitre le rendement et la qualité du fruit. On suppose qu'un des premiers caractères à avoir été sélectionné est l'hermaphrodisme permettant d'augmenter significativement le rendement et la régularité de production. D'autres caractères comme la vigueur, la tolérance aux pathogènes, la taille ou la couleur des baies, etc…, ont été au fil des générations sélectionnés par l'Homme. Les meilleurs cépages sélectionnés ont été par la suite maintenus et conservés par reproduction végétative.

Tout au long de leur Histoire, la vigne et le vin ont suscité de fortes considérations symboliques et religieuses. Dans l'Egypte ancienne, la vigne et le vin auraient été apportés par Osiris, chez les Grecs par Dionysos et chez les Romains par Bacchus (Figure 9). C'est ensuite la religion Chrétienne, en associant le vin au sang du Christ, qui sera très impliquée dans la diffusion de la viticulture. Aujourd'hui, vin et viticulture restent chargés de nombreux symboles sociaux, culturels et gastronomiques tout particulièrement en France (Saint-Emilion, inscrit au patrimoine Mondial de L'UNESCO en 1999).

1.6-Bouleversements modernes du vignoble

L'arrivée de l'oïdium puis du phylloxéra et du mildiou dans les années 1850-1870 en provenance des Etats-Unis, a modifié durablement la viticulture européenne (Legros, 1997). Le phylloxera a détruit une grande partie du vignoble européen, et celui-ci a été reconstruit sur la base de quelques cépages phares qui ont de ce fait supplantés les anciens cépages aujourd'hui présents presque exclusivement dans les collections. Le phylloxera est un puceron qui attaque principalement les racines de la vigne (Figure 10). Les espèces américaines ayant co-évolué en présence du puceron sont résistantes ou tolérantes à ses attaques. L'importation des vignes américaines en Europe a provoqué la dissémination du phylloxera à la vigne européenne sensible, entraînant jusqu'à 50% de perte du vignoble français en 1876 (Legros, 1993). La viticulture et les cépages européens auraient alors pu disparaitre.

Pour lutter contre le phylloxera, deux approches ont été suivies. La première méthode a consisté au greffage des cépages européens sur des plants résistants d'origine américaine. Cela a eu pour avantage de combiner la qualité de récolte des cépages traditionnels européens à la résistance racinaire apportée par le porte-greffe américain (Pouget, 1990). La deuxième approche consiste à effectuer des croisements inter-spécifiques entre espèces américaines et européennes (hybrides). Elle a permis de fournir du matériel qui était partiellement résistant

aux différentes maladies (oïdium, mildiou et phylloxera). Cependant la qualité de la récolte était généralement médiocre (Pouget, 1990). Les scandales des vins frelatés dans les années 1900-1930 provenant des hybrides ont entaché également fortement la réputation de ces derniers avec pour conséquence, en 1934, l'interdiction de six hybrides producteurs directs (Noah, Isabelle, Clinton, Herbemont, Othlello et Jacquez). Par la suite, sous l'influence des politiques viticoles nationales et européennes, les surfaces cultivées d'hybrides ont fortement régressé depuis 1958. Cependant une vingtaine d'hybrides restent inscrits au catalogue officiel français (Boursiquot *et al.*, 2007).

Depuis les années 1940, du fait de la médiocre qualité des hybrides producteurs directs et de la réputation des cépages déjà implantés en Europe, la création variétale est restée limitée en France. La création des Appellations d'Origine Contrôlée des vins et des eaux de vie (A.O.C.) en 1935 par Pierre Le Roy de Boiseaumarié et Joseph Capus, génère un nouveau frein à la création variétale (Doré & Varoquaux, 2006). Les domaines viticoles en A.O.C. s'emploient à respecter un cahier des charges strict. Entre autres, seuls certains cépages dont la liste est établie par l'INAO sont autorisés à la plantation sur un terroir précis. Aujourd'hui, 54% des vins français sont produits par des Appellations d'Origine Protégée (A.O.P., similaire à l'A.O.C mais au niveau Européen), 34% en vin de pays (IGP, Indication Géographique Protégée) et enfin 12% en vin de table (Statistique France AgriMer 2007). De ce fait, les efforts des sélectionneurs français se sont davantage portés dans la seconde moitié du XXème siècle sur les programmes d'amélioration des clones, des porte-greffes, des variétés de raisins de table et de cépages de cuve pour les vins de table (1947 clones agréés depuis 1971, (Yobregat *et al.*, 2011)).

Cependant avec la nécessaire baisse des intrants phytosanitaires (plan Eco-Phyto 2018) en lien avec le développement d'une agriculture durable et les changements climatiques, la création de nouveaux cépages est relancée en France depuis une dizaine d'années, tout comme en Europe, où de nombreux programmes de création variétale se sont poursuivis, comme en Allemagne ou en Hongrie, en particulier sur des cépages résistants au mildiou et/ou à l'oïdium (Chabin *et al.*, 2008).

1.7-Diversité de Vitis vinifera subsp. vinifera

La diversité de la vigne cultivée est très importante que ce soit au niveau des cépages ou des clones. Cependant, la variation entre les cépages (inter-variétale) est beaucoup plus

importante que la variation entre clones (intra-variétale) tant sur les caractères agro-morphiques que sur la diversité moléculaire (Boursiquot *et al*., 2007). Dans l'étude de Laucou *et al.* (2011), la diversité des cépages est comparée à celle des clones à partir de 20 marqueurs SSRs. De nombreux allèles ont été identifiés (6,9 allèles par locus SSR) permettant de différencier chacun des cépages selon leur profil SSR. Au contraire, chez les 97 clones analysés, 95% d'entre eux se sont avérés identiques attestant que la diversité intra-variétale est beaucoup plus faible que la diversité inter-variétale (dans les 5% restant, au maximum 4 loci se sont révélés polymorphes entre clones).

1.7.1-Les cépages

En 1956, Levadoux définit le terme cépage comme l'ensemble des individus possédant des caractéristiques ampélographiques très proches. « *Les cépages […] se présentent à nous comme des collections plus ou moins riches de clones plus ou moins voisins les uns des autres »*. En 2000, Boursiquot et This définissent le terme cépage comme un individu issu de reproduction sexuée ayant donné par la suite d'autres individus par multiplication végétative. Ils décrivent ainsi le cépage comme « *Une unité taxonomique propre à Vitis vinifera* L. *et qui est le produit d'un semis ou d'un individu unique au départ, multiplié par voie végétative ».* Par analogie, pour la plupart des autres plantes, le terme le plus proche de cépage est la variété, au sens de variété cultivée ou cultivar.

Les cépages se distinguent les uns des autres par leurs caractères ampélographiques, c'est-à-dire leurs différences phénotypiques. Seul le recoupement de plusieurs caractères permet d'identifier un cépage (Galet, 1985). Les caractères à étudier sont nombreux : la forme et la couleur des différents organes (extrémité du rameau en croissance, jeunes feuilles, feuilles adultes, rameaux, baies), la présence de poils (villosité), la texture des feuilles, etc… (Boursiquot & This, 1996).

1.7.2-Diversité des cépages

La vigne cultivée compte actuellement de 6000 (Galet, 2000) à 10 000 cépages (www.vitaceae.org) à travers le monde. Cependant 20 cépages représentent à eux seuls plus du tiers (37%) des vignobles mondiaux (Boursiquot J.M., Com. Pers.). La majorité de la diversité des cépages se retrouve donc cantonnée à des petites parcelles ou dans les conservatoires.

Figure 11 : Aperçu de la diversité des cépages dans la collection internationale du Domaine de Vassal (JM Boursiquot).

Durant le XIXème siècle, plusieurs événements tels que les introductions de maladies comme l'oïdium, le mildiou et phylloxéra ont conduit à la réduction de la diversité de l'encépagement (Cf. Chapitre1, Section-1.6). Ces crises ont entrainé une perte importante de la diversité génétique, puisque seuls les cépages les plus importants économiquement ont été replantés. Les anciens cépages ne sont aujourd'hui présents que dans les collections. Le passage de plantation pluri-cépage à mono-cépage et mono-clonal, ainsi que l'adoption des A.O.C., ont participé également à la diminution de la diversité d'encépagement dans les vignobles.

Pour préserver la diversité des cépages et conserver les ressources génétiques de cette espèce, des collections ont été mises en place (Bouquet & Boursiquot, 1999). La plus importante dans le monde se situe au Domaine de Vassal (INRA). Elle a été créée en 1876 à l'Ecole d'Agronomie de Montpellier puis déplacée sur le Domaine de Vassal en 1949 (Lacombe, 2009). Les cépages sont plantés sur une lagune sableuse les préservant des attaques du phylloxéra et des contaminations par le virus du court-noué, du fait de l'absence du nématode vecteur. C'est la collection de référence au niveau national et international. Elle comporte plus de 5500 accessions de *Vitis vinifera* L. correspondant à 2323 cépages uniques identifiés sur la base de marqueurs microsatellites (Laucou *et al.*, 2011). Cette collection présente une diversité morphologique considérable (Boursiquot *et al.*, 1995) et la figure 11 illustre une partie de cette diversité. La conservation de ces cépages est essentielle, pour des raisons patrimoniales mais également pour la création de nouvelles variétés.

1.7.3-Les clones

La définition précise des clones est récente : «*Un clone est la descendance végétative conforme à une souche choisie pour son identité indiscutable, ses caractères phénotypiques et son état sanitaire »* (OIV, 1996). On définit ainsi les clones comme des individus issus de multiplication végétative à partir d'un individu unique au départ. Ils sont donc issus du zygote ancestral unique d'un cépage qui a été reproduit par voie végétative. Ils sont définis par le même nom que le cépage dont ils sont issus, et lorsqu'ils sont officiellement agréés (section vigne du CTPS) d'un numéro spécifique (ex : Carignan, clone n°6).

Au sein d'un cépage, les clones possédant des caractéristiques ampélographiques communes sont à l'heure actuelle définis comme des variétés. Ainsi, le cépage « Pinot », regroupe toutes les variétés de Pinot : Pinot noir (clones de Pinot aux baies noires), Pinot

Varietal group	*Varieties or variants*	*Phenotype*	*Clonality ascertaining*	*References*
Cabernet Sauvignon	Cabernet Sauvignon	Black-skinned berries		
	Malian	Bronze-skinned berries	Bud sport of Cabernet Sauvignon	Boss *et al.*, 1996; Walker *et al.*, 2006
	Shalidan	White-skinned berries	Bud sport of Malian	
Chardonnay	Chardonnay blanc	White-skinned berries/ neutral aroma		
	Chardonnay muscaté	Rosy-skinned berries/ aromatic	16 SSR	Duchêne *et al.*, 2009
	Chardonnay rose	Rosy-skinned berries/ neutral aroma	Bud sport of Chardonnay blanc	This *et al.*, 2007
Italia	Italia	Green-skinned berries		
	Ruby Okuyama	Light-rosy-skinned berries	Bud sport of Italia	Kobayashi *et al.*, 2004
	Rubi	Light-rosy-skinned berries	Bud sport of Italia	Oliveira Collet *et al.*, 2005
	Benitaka	Rosy-skinned berries	Bud sport of Italia	Azuma *et al.*, 2009
	Brasil	Black-skinned and red-fleshed berries	Bud sport of Benitaka	Oliveira Collet *et al.*, 2005
Muscat d'Alexandria	Muscat d'Alexandria	White-skinned berries		
	Flame Muscat	Red-skinned berries	Bud sport of Muscat d'Alexandria	Kobayashi *et al.*, 2004
Pinot	Pinot noir	Glabrous leaves/ black waxed berries		
	Pinot moure	No wax berries	50 SSR	Hocquigny *et al.*, 2004
	Pinot meunier	Hairy leaves	50 SSR	Franks *et al.*, 2002
	Pinot gris	Red-grey-skinned berries	50 SSR	Hocquigny *et al.*, 2004; Walker *et al.*, 2006; Furiya *et al.*, 2009
	Pinot blanc	White-skinned berries	50 SSR/bud sport of Pinot gris	Hocquigny *et al.*, 2004; Walker *et al.*, 2006; Yakushiji *et al.*, 2006
Savagnins	Savagnins blanc	White-skinned berries/ neutral aroma		
	Savagnins rose	Rosy-skinned berries/ neutral aroma	16 SSR	Duchêne *et al.*, 2009
	Gewurztraminer	Rosy-skinned berries/ aromatic	16 SSR	
Ugni blanc	Ugni blanc	Fleshy berries		
	Fleshless mutant	Fleshless berries	Bud sport of Ugni blanc	Fernandez *et al.*, 2006

Table 1 : Exemples de quelques variétés pour lesquelles des clones ont été analysés par marqueurs moléculaires (D'après Pelsy, 2009).

blanc (clones de Pinot aux baies blanches), Pinot gris (clones de Pinot aux baies rose-gris) et Meunier (clones de Pinot fortement cotonneux) (Figure 1).

1.7.4-Diversité des clones

A l'instar d'autres plantes d'intérêt agronomique comme le citron ou le café, la diversité clonale chez la vigne est très importante (McKey *et al.*, 2009). En France, plus de 15 000 clones concernant 113 cépages sont répertoriés dans les conservatoires (Yobregat *et al.*, 2011). Parmi ces clones, 1163 sont aujourd'hui agréés et conservés dans la collection de référence de l'Institut Français de la Vigne et du Vin (IFV) au Domaine de l'Espiguette (Le Grau du Roi). Les clones non agréés dans un bon état sanitaire et présentant des caractéristiques agronomiques intéressantes, sont conservés dans les conservatoires régionaux.

Bien que conservant les caractéristiques variétales générales, les clones peuvent se distinguer les uns des autres par des caractères ampélographiques spécifiques. Certains se différencient par la découpure de leurs feuilles, la villosité, la pigmentation ou le port des rameaux etc... (Table 1 ; Pelsy, 2009). Les différences entre les clones peuvent porter sur des caractères quantitatifs comme par exemple la vigueur, la taille des baies, la teneur en sucre ou encore des variations d'arômes. Par exemple, le clone Chardonnay n° 76 est qualifié de neutre alors que le clone Chardonnay n° 809 présente des arômes muscatés. Ceux-ci sont dus à la forte quantité de linalol et géraniol produit par ce clone (Duchene *et al.*, 2009). D'autres clones possèdent des différences phénotypiques sur un caractère majeur : par exemple des différences dans la couleur des baies (Grenache blanc, gris et noir (Boursiquot *et al.*, 2007)), et sont alors considérés comme des variétés.

1.7.5-Etudes actuelles sur la diversité clonale

Les études réalisées ont toutes eu pour objectif l'identification des clones avec l'utilisation des marqueurs moléculaires afin de mieux appréhender la diversité clonale de la vigne et pour la future mise en place de tests commerciaux. Différents types de marqueurs moléculaires ont été utilisés sur différents échantillons avec plus ou moins de succès. En l'état des connaissances actuelles (500 SSRs disponibles dans la base de donnée de NCBI www.ncbi.nlm.nih.gov) les marqueurs de type SSRs, utilisés dans plusieurs études sur différents échantillons, ne sont pas suffisamment polymorphes pour permettre l'identification

d'un profil unique pour chaque clone (Riaz *et al.*, 2002; Hocquigny *et al.*, 2004; Moncada *et al.*, 2006; Moncada & Hinrichsen, 2007; Pelsy *et al.*, 2010). Une étude comparative sur la capacité des différents types de marqueurs moléculaires à discriminer les clones a été réalisée par Imazio *et al.* (2002). La conclusion est que les SSRs ne sont pas assez polymorphes alors que les marqueurs révélant du polymorphisme d'insertion tels que les AFLPs (basés sur la présence ou l'absence de fragments d'ADN) et S-SAP (basés sur la présence ou l'absence d'éléments mobiles) semblent au contraire, suffisamment résolutifs. De récentes études basées sur les AFLPs ont montré un niveau de polymorphisme permettant une identification plus efficace des clones (Konradi *et al.*, 2007; Anhalt *et al.*, 2010; Meneghetti *et al.*, 2011). Les marqueurs de type S-SAP semblent aussi donner de bons résultats pour identifier les clones (Wegscheider *et al.*, 2009). Cependant des études faites à partir d'amorces spécifiques à *Vine-1* de la famille Gypsy (Labra *et al.*, 2004) ou des amorces non-spécifiques de la famille Gypsy (Pereira *et al.*, 2005) n'ont pas montré de polymorphisme clonal. Il semble donc que la définition des amorces et le choix de l'élément mobile analysé soient très importants pour observer le polymorphisme entre les clones.

1.8-La sélection clonale en France

La sélection clonale a connu un essor très important depuis les années 1960 (Grenan *et al.*, 2000). C'est en effet jusqu'à présent, le seul moyen pour les viticulteurs de se procurer du matériel innovant tout en conservant l'identité et la typicité d'un cépage. Le premier objectif de la sélection clonale a été essentiellement sanitaire. En effet, les viroses dont principalement le court-noué et l'enroulement, sont graves et incurables. Dans un second temps, durant les dernières décennies, la sélection s'est aussi attachée à mettre à disposition de la filière, des clones aux caractéristiques agronomiques et technologiques, en adéquation avec les besoins des professionnels. Désormais, les besoins sont plus spécifiques : grappes plus lâches avec des baies relativement petites, port plus érigé facilitant le palissage, ou encore arômes particuliers. Ainsi, ont été récemment agréés en France deux clones de Gamay à port érigé (n°1108 et n°1109), un clone de Roussanne plus fertile (n°1040), deux clones de Viognier, à grappes plus lâches (n°1042) ou moins fertile (n°1051), un clone de Chardonnay très peu fertile et à petites grappes lâches (n°1066).

Les programmes de sélection clonale sont coordonnés par l'IFV et menés en collaboration avec les Partenaires de la Sélection Vigne, composés principalement des chambres

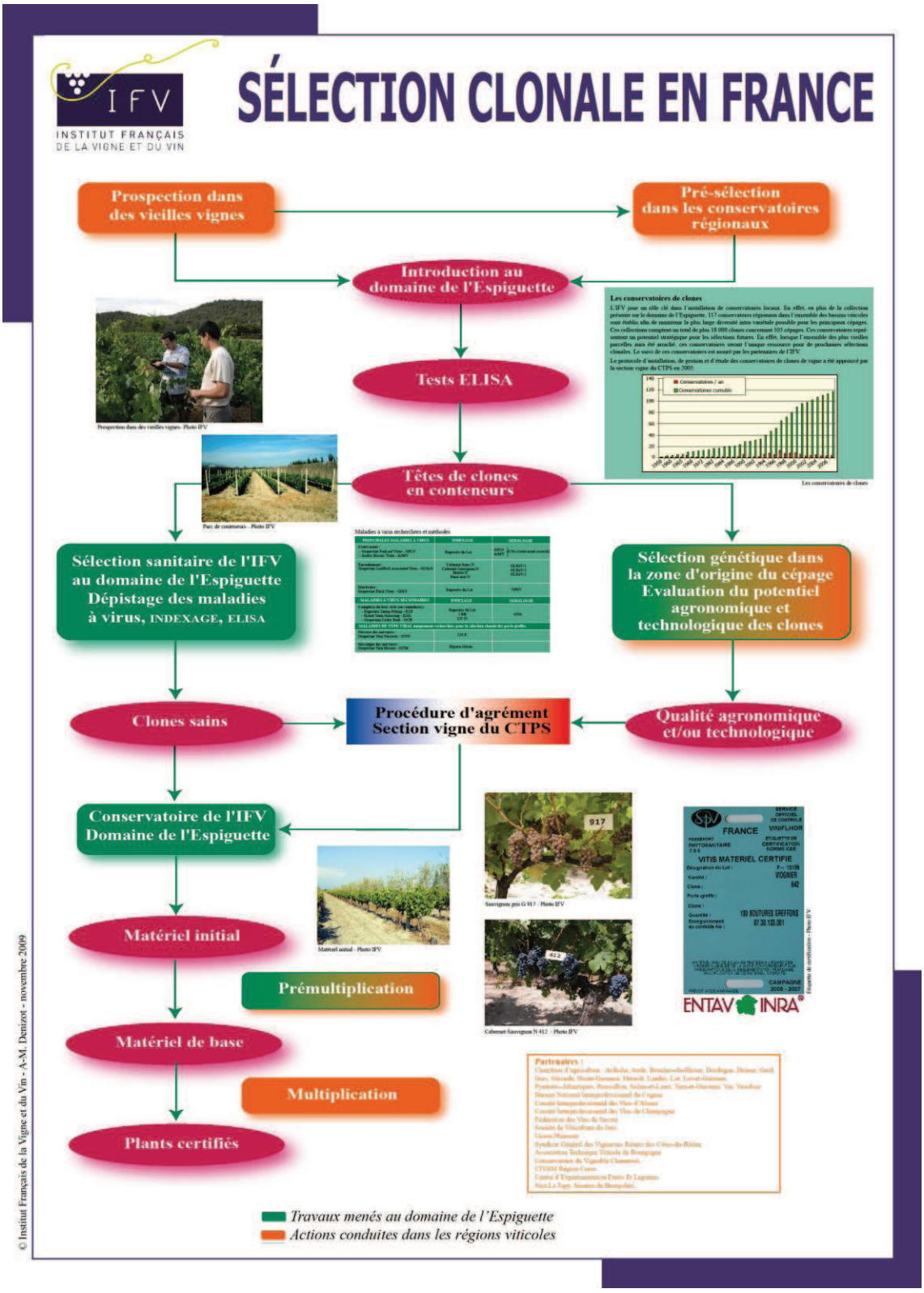


Figure 12 : Présentation du programme de sélection clonale mené par l'IFV en France.

d'agriculture, comités interprofessionnels, syndicats et de quelques associations (Figure 12 ; Boursiquot *et al*., 2007). Dans un premier temps, les clones d'intérêts sont sélectionnés durant des prospections dans les différents vignobles pour être plantés et évalués dans les conservatoires régionaux ou des parcelles expérimentales. Les clones possédant des caractères ampélographiques d'intérêts, sont introduits au Domaine de l'Espiguette où des tests sanitaires sont réalisés (Indexages et/ou tests Elisa pour le court-noué, enroulement, marbrure…), (Walter & Martelli, 1998). Les clones sont ensuite évalués sur leurs potentiels agronomique et technologique (qualité du vin produit). Une fois tous les tests effectués avec succès, le clone peut être présenté par l'IFV ou l'INRA (les deux centres de sélections officiels en France) à l'agrément à la section vigne du CTPS (Comité Technique Permanent de la Sélection des plantes cultivées). Le clone est ensuite planté dans le conservatoire du Domaine de l'Espiguette (matériel initial) où il sera multiplié pour, *in fine*, après plusieurs étapes, être produit par les pépiniéristes et vendu aux viticulteurs en catégorie certifiée. Depuis la prospection jusqu'à l'agrément et la multiplication dans les pépinières, il se déroule en moyenne 15 ans avant qu'un clone ne soit commercialisé et diffusé. Tous les clones étudiés au cours de cette thèse proviennent des programmes de sélection clonale ENTAV-INRA®.

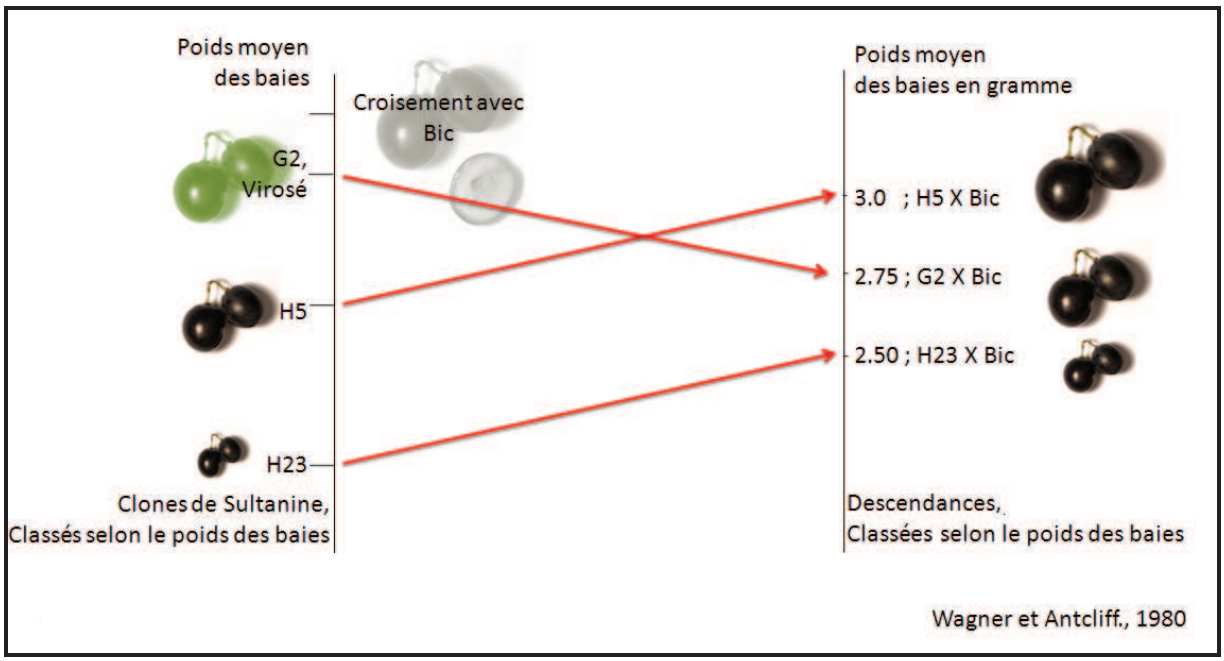


Figure 13 : Expérience de Wagner et Antcliff (1980) sur l'héritabilité de la variation clonale.

2-La variation clonale chez la vigne

2.1-Hypothèses sur l'origine de la variation clonale

Plusieurs hypothèses ont été proposées pour expliciter la variabilité clonale observée. Nous les présentons ici en deux groupes : les variations dues à l'état sanitaire de la plante et les variations dues à l'état et la structure du génome.

2.1.1-Variations dues aux virus

Une des hypothèses envisagées pour expliquer la variation clonale était une origine sanitaire. Cette dernière a été étudiée par Wagner & Antcliff, (1980). Dans leur expérience, ils disposaient de trois clones: deux sains et un virosé, possédant des baies de différentes tailles. Ces clones ont été croisés avec un même individu d'une autre variété possédant de grandes baies (Figure 13). Les descendants des clones sains présentaient une taille moyenne des baies supérieure aux clones parentaux. A contrario, la descendance du clone virosé a présenté une taille moyenne des baies inférieure à celle du clone intial. Les auteurs en ont donc conclu que la variation générée par le virus n'était pas héréditaire au contraire des mutations somatiques qui peuvent être transmises à la descendance. Depuis, aucune autre étude n'a montré l'influence de virus affectant de façon héréditaire un individu (Walter & Martelli, 1998; Mannini *et al.*, 1999). Les programmes de sélection clonale garantissent un bon état sanitaire permetant de se limité aux caractéristiques agronomiques intrinsèques de l'individu.

Figure 14 : Perte sectorielle de la chlorophylle dans une feuille de vigne (Photographie B. Molot).

2.1.2-Les variations dues à des modifications de la structure ou de l'état du génome

Aujourd'hui, sous réserve de considérer des individus adultes et avec un état sanitaire comparable, l'accumulation de mutations somatiques (génétiques ou épigénétiques) est l'hypothèse la plus parcimonieuse pouvant expliquer la diversité clonale (Hartmann *et al.*, 2001).

Bien que la multiplication végétative assure en théorie le maintien de l'identité génétique de la plante mère à la plante fille, des mutations somatiques peuvent apparaitre et être transmises à la descendance (Pelsy, 2009). Lors de leur développement, les individus accumulent ces mutations somatiques. La plupart de ces mutations sont silencieuses, c'est-à-dire qu'elles n'affectent pas le phénotype. Cependant, dans de rares occasions, elles peuvent aussi entrainer une modification de la fonction ou de la régulation d'un ou plusieurs gènes et avoir ainsi des conséquences sur le phénotype ; on parle alors de mutations non silencieuses (Sunyaev *et al.*, 2003). Ces mutations affectant le phénotype sont à la source de la diversité et de l'adaptation des espèces à leur milieu qui est en perpétuel changement (Darwin, 1859).

Des variations entre clones peuvent être aussi dues à la persistance ou à la disparition de caractères juvéniles. Lors de leur développement, les premiers organes mis en place par la vigne possèdent en effet des caractères dits juvéniles. Par exemple, la forme et la couleur des premières feuilles sont différentes des feuilles à l'état adulte (Martinez *et al.*, 1997). Lors de la multiplication de clones après passage en culture *in vitro*, on a constaté que selon le stade de développement et l'organe utilisé lors du bouturage, la durée du stade juvénile pouvait varier (Grenan & Truel, 1983).

Une illustration d'une mutation somatique spectaculaire connue chez les plantes est la perte de chlorophylle (Gustafsson, 1947). Si une cellule subit une mutation touchant une enzyme de la voie de biosynthèse de la chlorophylle, celle-ci perd alors la capacité à produire ses pigments chlorophylliens. Cette mutation peut être propagée à toute une partie d'un organe par divisions mitotiques de la cellule mutante, comme par exemple cette mutation sectorielle de feuille de vigne (Figure 14). Dans le cas où une mutation somatique apparait dans toutes les cellules initiales d'un bourgeon et que celui-ci est sélectionné pour produire un nouveau plant, alors cet individu sera porteur de cette mutation et pourra donner naissance à

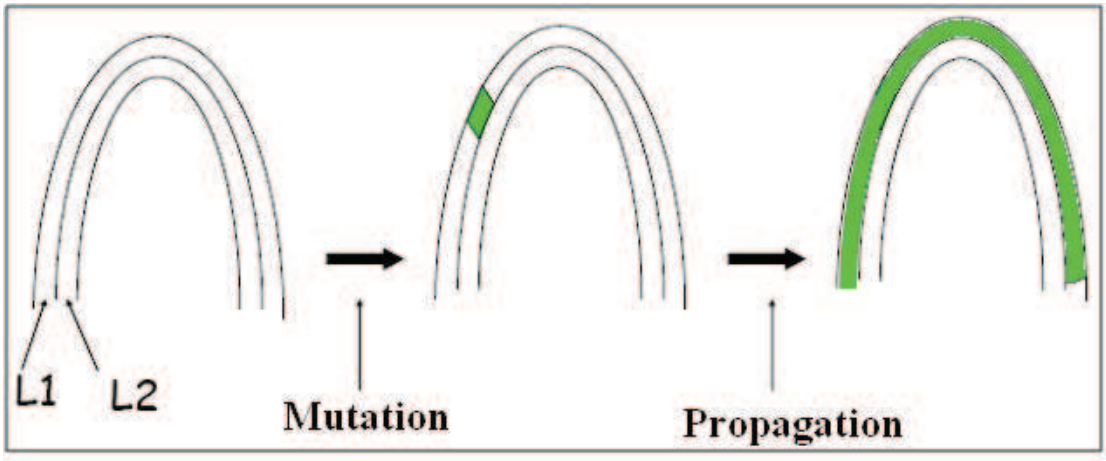


Figure 15 : Description schématique de l'apparition et de la propagation d'une mutation chimérique.

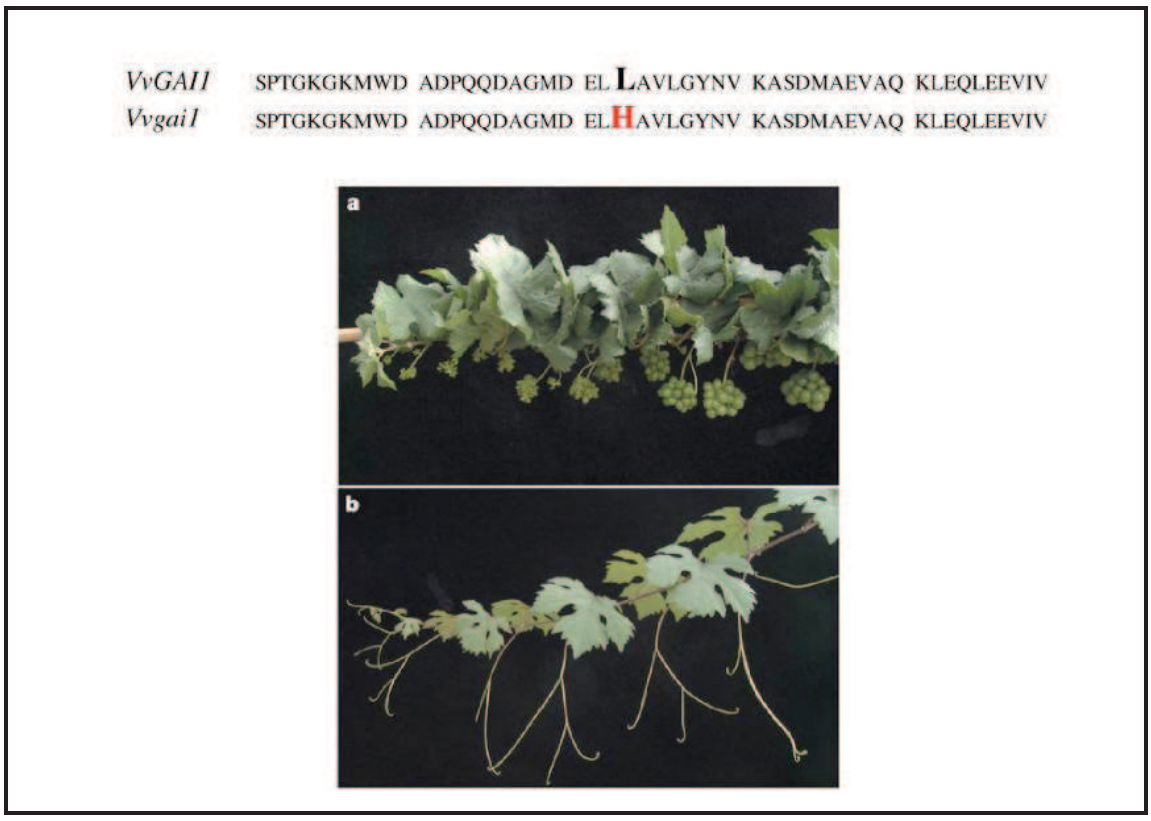


Figure 16 : La mutation chimérique du Meunier. Mutation sur le gène *VvGAI1* contenu dans la couche L1 (L versus H). Les individus issus de la régénération après embryogénèse somatique présentent a) un phénotype nain si issus de la couche L1 ou b) un phénotype sauvage si issus de la couche L2 (Boss *et al*., 2002).

un nouveau clone. De plus, la vigne, comme la plupart des plantes, est composée d'au moins deux lignées cellulaires (L1 et L2). La lignée cellulaire L1 donne naissance aux tissus épidermiques et la lignée L2 donne naissance aux tissus vasculaires et parenchymateux ainsi qu'aux gamètes. Dans certains cas, une mutation apparait dans une seule lignée cellulaire (L1 ou L2) on parle alors de mutation chimérique (Figure 15). Ce type de mutation est assez couramment observé chez la vigne (Franks *et al.*, 2002). A titre d'exemple, le Meunier possède une mutation ponctuelle sur le gène *VvGAI-1* localisé sur la couche cellulaire L1. Cette mutation rend insensible un récepteur aux hormones gibbérellines et provoque le phénotype nain chez les individus régénérés par embryogénèse somatique à partir de cette couche cellulaire. Au contraire, les plantes régénérées à partir de la couche L2 présentent, elles, un phénotype normal (Figure 16 ; Boss & Thomas, 2002).

2.2-Différents types de mutations somatiques

Les mutations somatiques peuvent être d'origine génétique, c'est-à-dire entrainant des variations dans la séquence de l'ADN, ou épigénétique et entrainant des variations au niveau de la structure de l'ADN. On trouve ainsi différents types de mutations génétiques : les mutations ponctuelles (SNP et indel), les éléments mobiles ou transposables et les variations structurales du génome.

La polyploïdisation, est également un phénomène pouvant perturber fortement le phénotype. Chez les plantes, la polyploïdisation est courante (colza, maïs, asperge) (Falavigna & Casali, 2002; Levy & Feldman, 2002). Chez la vigne, la polyploïdisation peut être naturelle ou provoquée *in vitro* pour générer des individus triploïdes, tetraploïdes ou hexaploïdes (Yamashita *et al.*, 1998). Ces individus sont reconnaissables (feuille épaisses, entre-nœuds courts) et ne sont généralement pas pris en compte dans de la cadre de la sélection clonale. Ils possèdent une fertilité limitée et ont eu finalement peu de succès (Park *et al.*, 1999).

2.2.1-Les mutations génétiques

2.2.1.1- Les événements ponctuels

Ce type de mutations regroupe les substitutions d'un acide nucléique par un autre (SNP, Single Nucleotide Polymorphism) et les insertions délétions (indel) (Antoni, 2002). Ce type de polymorphisme est certainement le plus étudié actuellement. Il est facilement détectable et

Echantillon	Nombre de gènes	SNP/ nombre de nucléotides séquencés	Référence
9 cultivés	230	1/64	Lijavetzky et al, 2007
92 cultivés	3	1/49	Le Cunff et al, 2007
24 cultivés	35	1/104	Vezzulli et al, 2008

Table 2 : Diversité génétique chez *Vitis vinifera* L. établie à l'aide des SNPs.

peut, dans certains cas, être directement associé à une différence phénotypique (Konishi *et al.*, 2006; Yeager *et al.*, 2007). La fréquence de ces mutations varie énormément selon leur nature de ces dernières, les organismes ainsi que l'environnement. De nombreuses molécules mutagènes peuvent augmenter très significativement le taux de ces mutations (ex EMS, utilisé pour produire des banques de mutants (Wu *et al.*, 2005)).

Les substitutions d'acides nucléiques

Les SNP sont des changements entre deux nucléotides (A-G ; T-C ; A-T ; C-G). Le taux de mutations des SNPs est en moyenne 2,5 x 10^{-8} nucléotides par génération (Nachman & Crowell, 2000) mais varie selon leur localisation dans le génome (quatre fois plus fort chez le chromosome X que le chromosome Y chez l'Homme. De façon générale, les régions soumises à sélection, telles que les exons des gènes, ont un taux de mutation plus faible que les régions inter-géniques. Chez la vigne, les SNPs ont été recherchés principalement dans les gènes (Table 2 ; Lijavetzky *et al.*, 2007; Le Cunff *et al.*, 2008; Vezzulli *et al.*, 2008). L'ensemble de ces études montrent que malgré la disparité du nombre de gènes étudiés et du nombre d'individus, la diversité génétique est importante chez *V. vinifera* L. avec un SNP toutes les 64 bases (Lijavetzky *et al.*, 2007) et jusqu'à 1 SNP toutes les 49 bases (Le Cunff *et al.*, 2008). La diversité nucléotidique de la vigne (nombre moyen de différences nucléotidiques par site entre paires de séquences prises au hasard) est de 0,0051 (Lijavetzky *et al.*, 2007) similaire à celle du maïs (0,0063 ; Ching *et al.*, 2002)) et environ 5 fois plus importante que celle du soja (0,0012 ; Zhu *et al.*, 2003)) et de la tomate (0,0010 ; Labate *et al.*, 2009)).

Les insertions et délétions

Les mutations de type insertions / délétions (indels) sont gain ou une perte d'une ou plusieurs bases dans la séquence originale. Ces indels peuvent apparaitre suite à une mauvaise réparation de l'ADN après un événement de recombinaison (Dong *et al.*, 2002), une insertion / excision d'un élément transposable (Yamashita *et al.*, 1999; Van de Lagemaat *et al.*, 2005) ou des duplications (Chen *et al.*, 2005). De la même façon que les SNPs, leur taux de mutation varie énormément selon leur localisation dans le génome, en particulier si les indels sont situés dans des motifs dit microsatellites (SSRs). Les microsatellites sont des motifs particuliers que l'on retrouve dans tous les génomes. Ils sont constitués par la répétition d'un motif d'une à six bases répété *n* fois. Le nombre de répétition varie énormément. La fréquence de mutation des SSRs par génération est élevée et évaluée à 10^{-4}

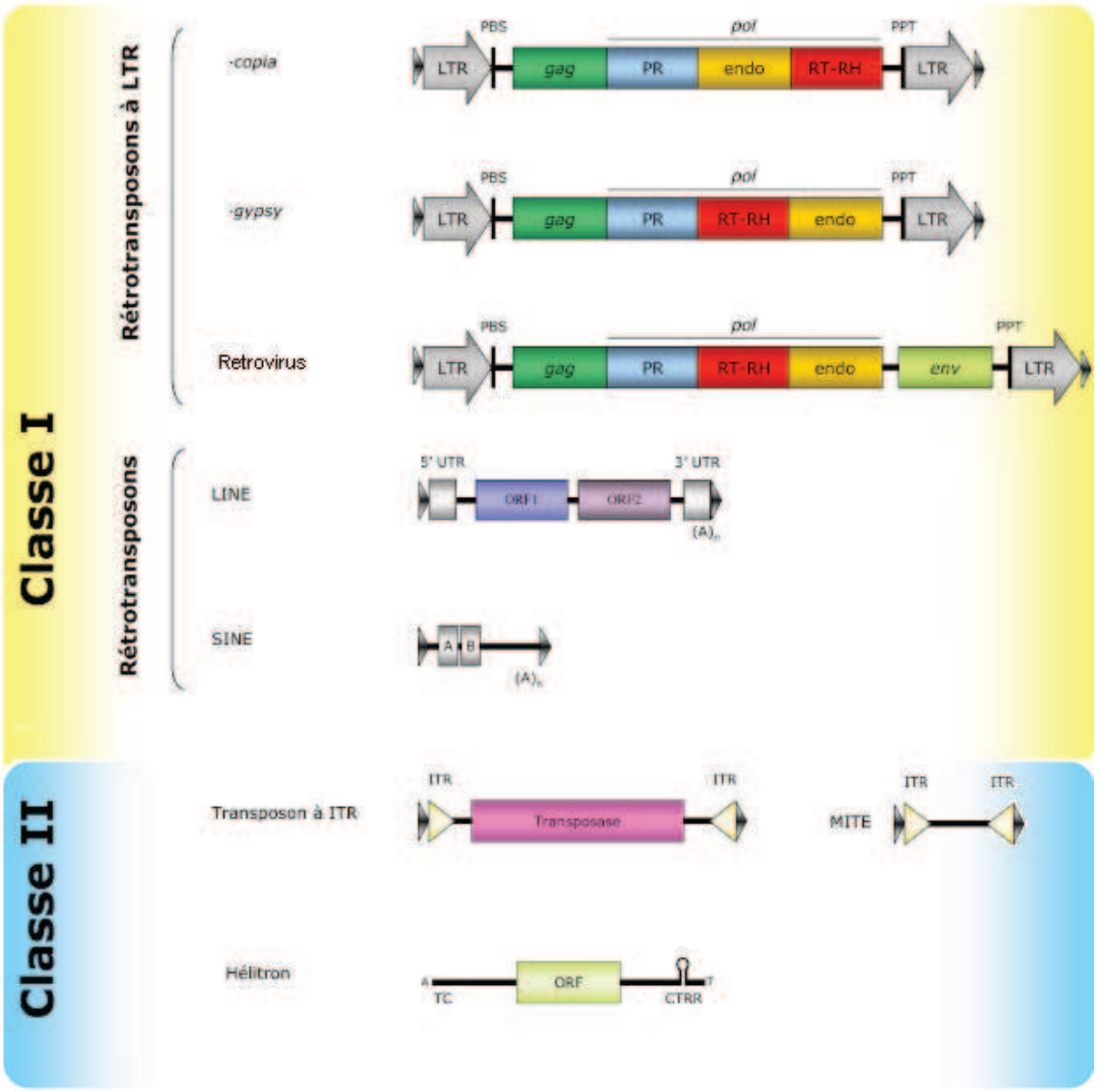


Figure 17 : Structure des éléments transposables connus chez la vigne (Moisy, 2008).

(Vigouroux *et al.*, 2002) du fait de la nature de la séquence le constituant. En effet, la répétition des motifs facilite les cassures de l'ADN et rend difficile la réparation de celui-ci (Li *et al.*, 2002). De par leur fort taux de polymorphisme, les SSRs ont été ces dernières années le type de marqueur moléculaire le plus utilisé. Chez la vigne, ils se sont révélés très efficaces pour analyser la diversité des populations (Aradhya *et al.*, 2003; Laucou *et al.*, 2011), pour établir des relations phylogénétiques (Péros *et al.*, 2010), pour construire des cartes génétiques (Doligez *et al.*, 2006), ou pour l'identification des cépages (This *et al.*, 2004).

2.2.1.2-Les éléments transposables

Les éléments mobiles ou éléments transposables ont été découverts par Barbara McClintock (1953) chez le maïs. Cependant, il a fallu plusieurs dizaines d'années pour que leur importance soit reconnue par la communauté scientifique (B. McClintock, Prix Nobel en 1983). Ces éléments sont constitués par une séquence d'ADN particulière capable de se déplacer ou de ce transposer dans le génome de la plante hôte. Les éléments transposables constituent la majeure partie de l'hétérochromatine et sont présents dans tous les organismes (McDonald, 1995). Au vu du nombre et de la diversité des éléments transposables, nous ne présenterons ici que les familles d'éléments mobiles connus chez la vigne.

Nomenclature et représentation des éléments transposables

Le débat qu'entraîne la problématique de la classification des éléments transposables n'est pas sans rappeler celui de la classification des organismes (Seberg & Petersen, 2009; Wicker *et al.*, 2009). La nomenclature utilisée dans ces travaux est basée sur celle proposée par Wicker *et al.* (2007) qui repose sur le mécanisme de transpostions et sur la structure des éléments mobiles. Au vu des données croissantes d'annotations et de séquençages des génomes et de la découverte de nouveaux mécanismes de transposition et de régulation, cette classification évoluera certainement.

Les éléments transposables sont répartis en deux grandes classes selon le mécanisme de transposition : la classe I ou rétrotransposons, qui présente un mécanisme de transposition de type « copier-coller » ; la classe II ou transposons, possédant un mécanisme de type « couper-coller ». Ces deux classes possèdent des éléments autonomes et non-autonomes, c'est-à-dire capables ou non de synthétiser toutes les enzymes nécessaires à leur réplication.

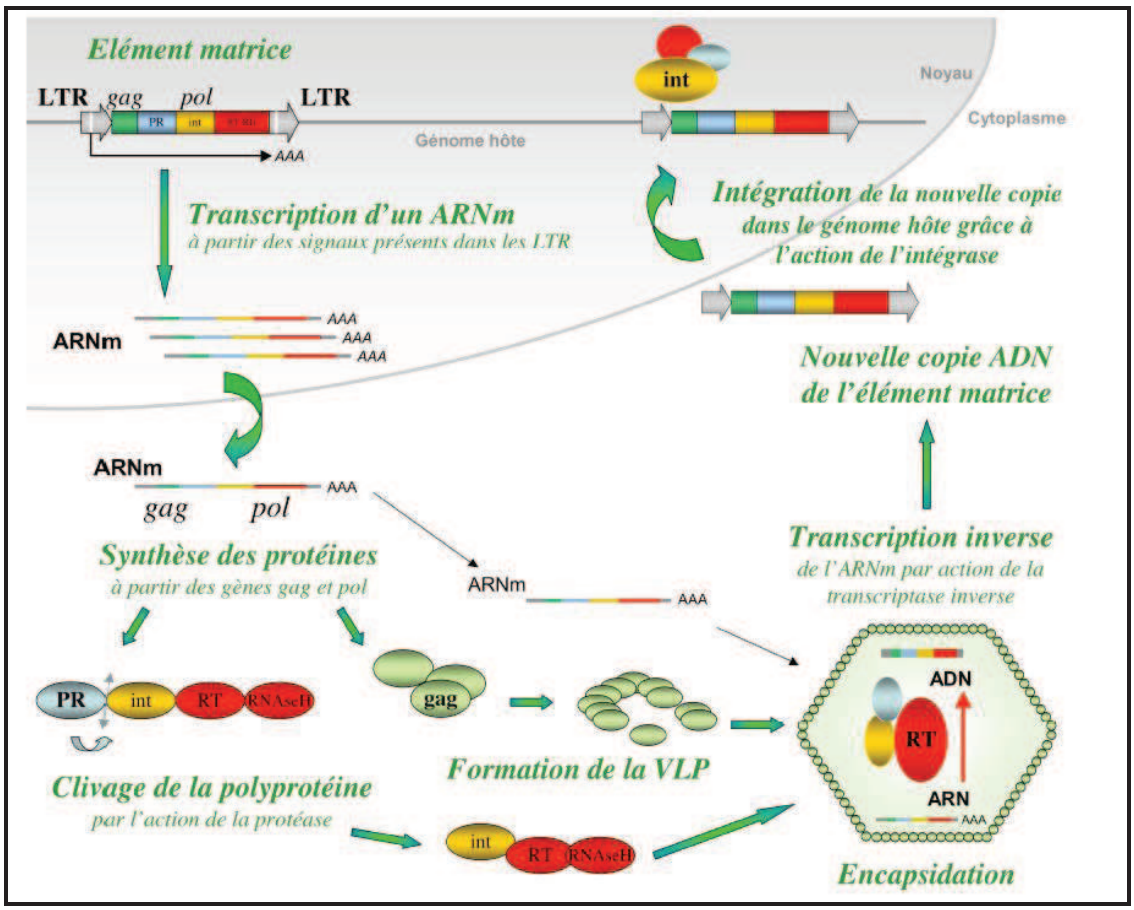


Figure 18 : Cycle de transposition des rétrotransposons à LTR (Moisy, 2008).

En 2007, l'annotation du génome de la vigne (Jaillon et *al*., 2007) a permis de dresser un panorama plus exhaustif des différentes familles d'éléments transposables présents chez la vigne. Au total, 111 879 éléments transposables ont pu être identifiés, représentant 17% du génome. Parmi eux, 12,4% se situent dans les introns des gènes annotés. La plupart des supers familles d'éléments mobiles sont représentées dans le génome de la vigne (Figure 17). Des études *in silico* sur la séquence du génome de la vigne ont permis de mettre en évidence de nouveaux éléments de classe II (Benjak *et al.*, 2008; Benjak *et al.*, 2009), ainsi que de classe I (Pelsy & Merdinoglu, 2002). La base de données d'éléments transposables de la vigne disponible sur RepBase (Jurka *et al*., 2005) comportait 107 éléments en 2007 et 191 en 2011.

Classe I, les rétrotransposons

Les éléments de classe I ont un mécanisme de transposition nécessitant la transcription d'une molécule intermédiaire d'ARNm (Figure 18). Cet ARNm est à la fois utilisé pour produire les protéines nécessaires à la transposition et utilisé également comme matrice pour produire de nouvelles copies de l'élément. Dans un premier temps, la transcription des gènes *gag* et *pol* permettent la mise en place respectivement d'une nucléo-capside et du complexe enzymatique nécessaire à la transposition. Une fois la nucléo-capside formée, l'ARNm va être encapsidé et transcrit en ADN grâce à l'enzyme Reverse Transcriptase (RT). Les nouvelles copies de l'élément vont ensuite être intégrées dans le génome grâce à l'enzyme intégrase. Ainsi, un événement de transposition peut générer plusieurs insertions de nouvelles copies de l'élément dans le génome.

Les rétrotransposons à LTR

Les rétrotransposons à LTR contiennent à leurs extrémités des séquences répétées orientées dans le même sens appelées Long Terminal Repeat (LTR). Celles-ci contiennent, entre autre, un site dit de l'intégrase, d'une vingtaine de bases permettant leur bonne intégration dans le génome. La séquence interne du rétrotransposon contient généralement les gènes *gag* et *pol* qui sont nécessaires à la transposition (Figure 17).

Parmi les différentes familles de rétrotransposons à LTR, les familles des *copia* et des *gypsy* sont très représentées dans le génome des plantes (Feschotte *et al.*, 2002). Ces éléments font en général plusieurs kilobases, jusqu'à 23 Kb pour *ogre* (Neumann *et al.*, 2003), et peuvent être autonomes ou non-autonomes (Wicker *et al*., 2007). De nombreux éléments de ces familles ont été identifiés dans le génome de la vigne et représentent au moins 14% de son

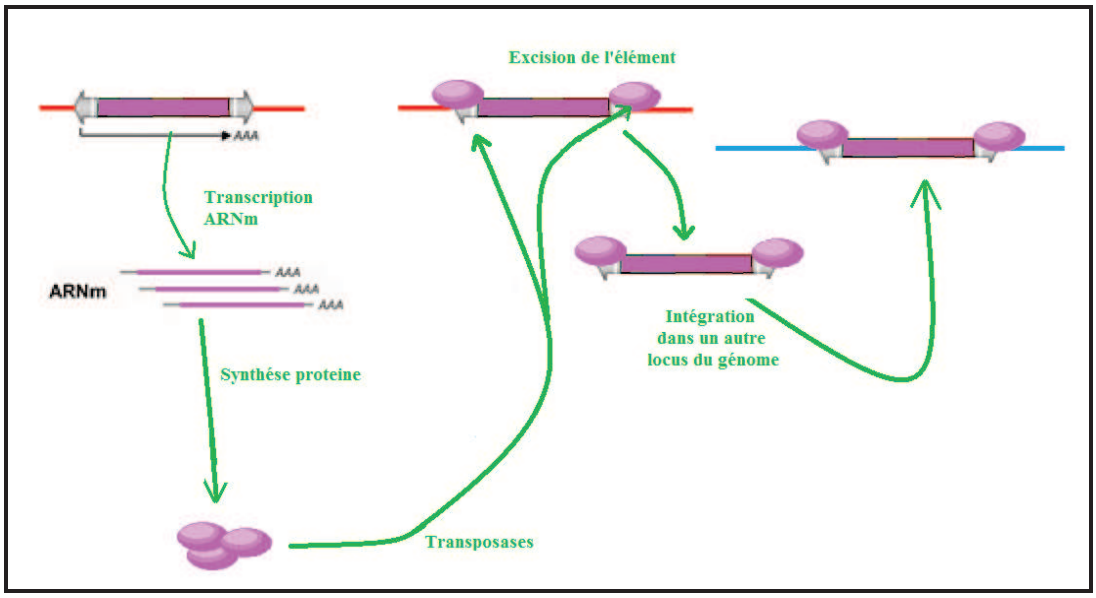


Figure 19 : Cycle de transposition des transposons.

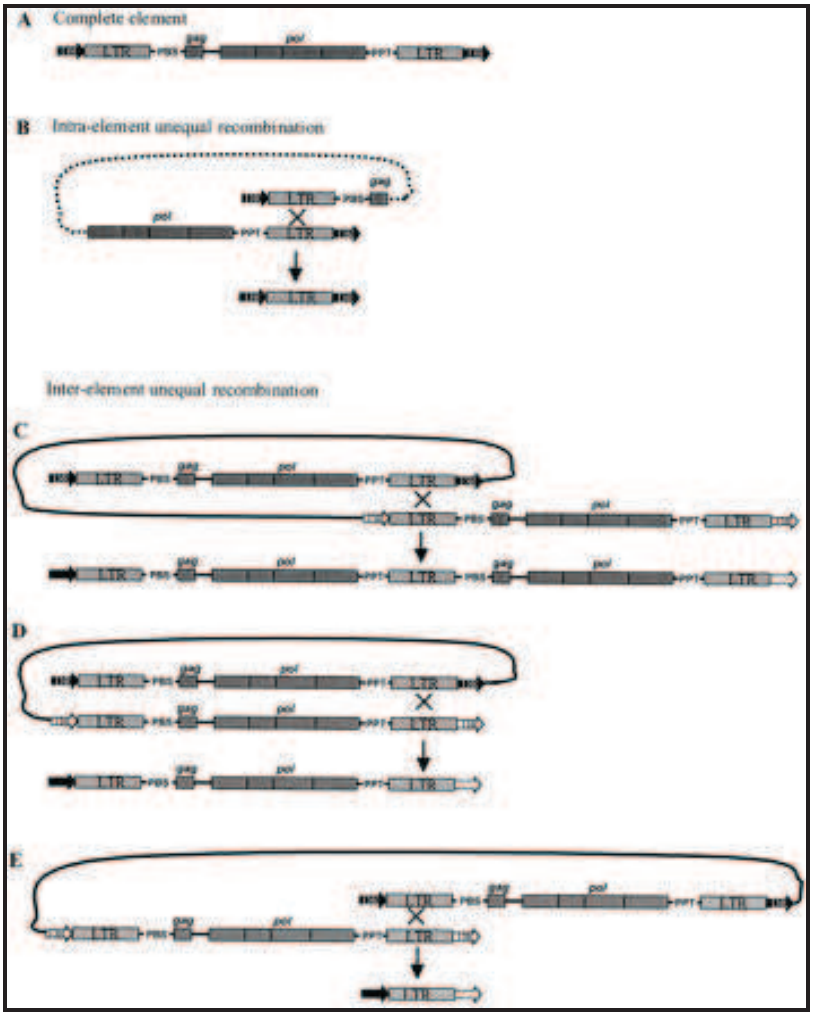


Figure 20 : Différents mécanismes d'élimination des éléments transposables (Devos *et al*., 2002).

génome (Jaillon *et al*., 2007). La famille des retrovirus sont des rétrotransposons à LTR particuliers, capables d'infecter les cellules voisines ainsi que d'autres individus hôtes (Bannert & Kurth, 2006). A l'heure actuelle 3 rétrovirus *caulimoviridae* ont été découverts chez la vigne *in silico* (Jaillon *et al.,* 2007).

Les LINEs, Long Interspersed Nuclear Elements

Ce sont des rétrotransposons sans LTR. Ces éléments mobiles sont autonomes, les gènes nécessaires à leur transposition sont présents dans leur séquence (Figure 17). Parmi les différentes familles qui existent, la famille L1 est la plus connue et étudiée. Les éléments L1 sont très présents chez les mammifères (20% du génome humain) mais peu représentés chez les insectes (3% pour le moustique). Chez les plantes, la fréquence des L1 semble très variable en fonction de l'espèce observée (6 copies chez *A. thaliana,* 100 chez *Z. mays* et 250 000 chez *L. speciosum* (Feschotte *et al*., 2002)). L'annotation du génome de référence de la vigne a permis d'établir que la famille L1 représente 75% de tous les éléments mobiles détectés chez la vigne (Jaillon *et al.* 2007).

Les SINEs, Short Interspersed Nuclear Elements

Ce sont des éléments de petites tailles (quelques dizaines ou centaines de bases) constitués d'une région régulatrice . Ils sont non-autonomes c'est-à-dire qu'ils ne sont pas capables de transposer sans l'aide des autres éléments. L'élément SINE le plus connu est *Alu,* qui représente 10% du génome humain (The International Human Genome Sequencing Consortium, 2001). L'annotation des SINEs est très difficile du fait de leur nombre, leur diversité et leur taille. Aucun SINE n'a été annoté jusqu'à présent dans le génome de la vigne.

Classe II, Les transposons et les hélitrons

Les éléments de classe II ont un mécanisme de transposition ne nécessitant pas la transcription d'une molécule intermédiaire d'ARNm. L'ARN produit ne sert qu'à synthétiser les protéines nécessaires à leur transposition (transposases). La transposition commence par l'excision de l'ADN de l'élément (Figure 19). Celui-ci va être alors pris en charge par les transposases pour être inséré dans un autre endroit du génome. Généralement, après avoir été excisée, la molécule d'ADN est amplifiée de telle façon que de multiples insertions de différentes copies de la molécule seront possibles dans le génome.

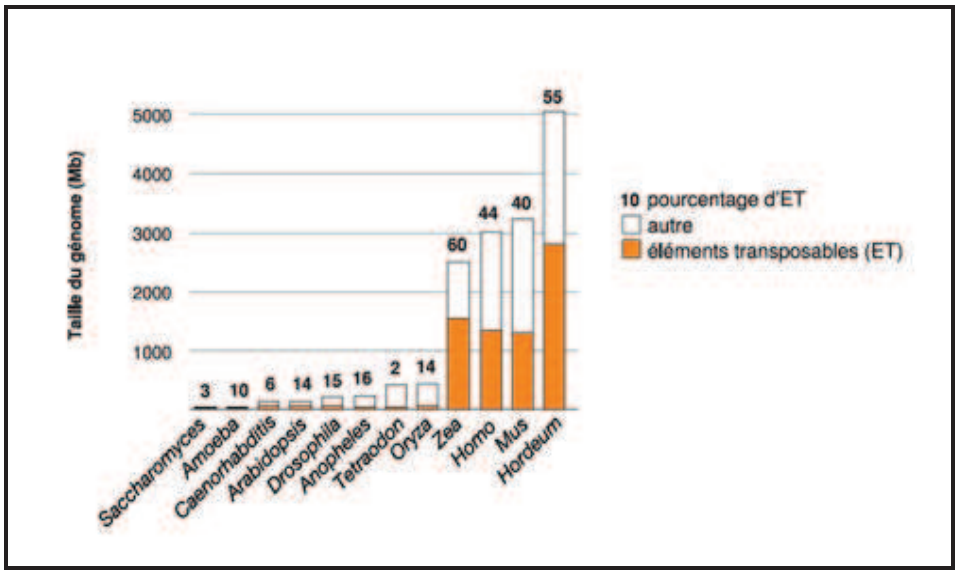


Figure 21 : Proportion d'éléments transposables en fonction de la taille des génomes (Kidwell *et al*., 2002).

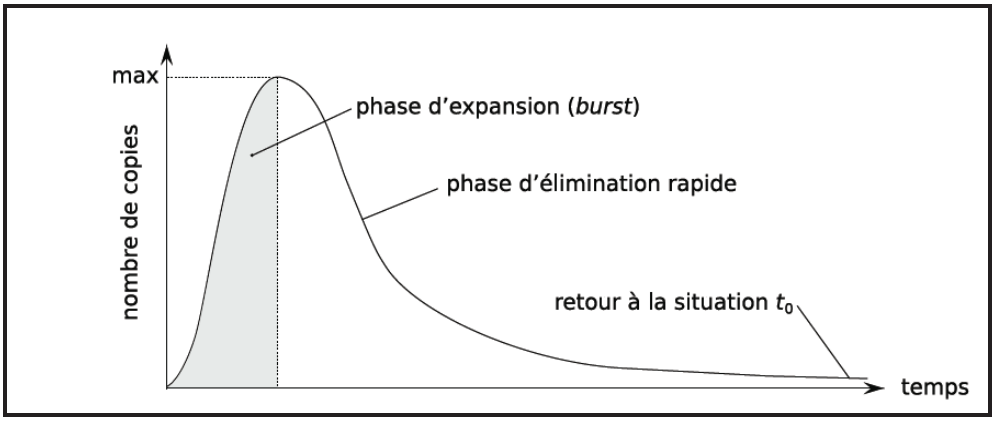


Figure 22 : Activité de transposition d'un élément transposable au cours de sa vie (Moisy, 2008).

Transposons à ITR

Ils sont constitués d'une séquence encadrée par des séquences répétées inversées (Inverted Terminal Repeats, ITR) et contiennent les gènes nécessaires à sa transposition (les transposases ; Figure 17). Ces éléments sont présents dans tous les génomes eucaryotes, leur taille dépassant rarement quelques kilobases. Parmi les familles de transposons connues, le génome de la vigne comporte des éléments de la famille des Mutator, CACTA et hAT.

Les MITEs (Miniature Inverted-repeat Transposable Element) sont des éléments non autonomes. Les MITEs regroupent les anciens transposons à ITR non autonomes, et ne forment pas, pour cette raison, une famille à part entière mais un sous-ensemble des transposons à ITR. Ils mesurent en général entre 500 et 1000 bases, et leurs transpositions nécessitent l'aide des transposons actifs. Chez la vigne, l'étude de Benjak *et al.* (2009) a permis d'identifier des éléments MITEs dérivant de toutes les familles de transposons.

Les hélitrons

Leur transposition ne fait pas appel à une transposase mais à un mécanisme appelé « Rolling-circle Replication » (Kapitonov & Jurka, 2001). Pour transposer, ces éléments utilisent les protéines de réplication de la cellule pour la synthèse de nouvelles copies. Ils ont été détectés dans le génome de nombreuses plantes, animaux et champignons. Chez la vigne un seul élément est actuellement connu : *hélitron-1* (Jaillon *et al.,*2007).

Régulation de l'activité des éléments mobiles

L'accumulation de mutations et les recombinaisons intra ou inter-éléments sont des mécanismes qui participent de manière importante à l'élimination ou à l'inactivation des éléments transposables. La recombinaison intra ou inter-éléments s'accompagne de la perte d'éléments transposables mais aussi de fragments d'ADN (Figure 20). En outre, la recombinaison n'élimine pas en général totalement l'élément transposable. Par exemple, la recombinaison entre LTR de rétrotransposons s'accompagne de la formation de solo-LTR qui est conservée dans le génome (Figure 20 ; Devos *et al.*, 2002). Ces éliminations peuvent avoir des conséquences sur les gènes. Par exemple, chez l'orge, la recombinaison entre les LTRs de deux copies différentes de l'élément *Bare-1* a provoqué la formation d'un solo-LTR et l'élimination d'une partie du gène *rar-1* (Shirasu *et al.*, 2000).

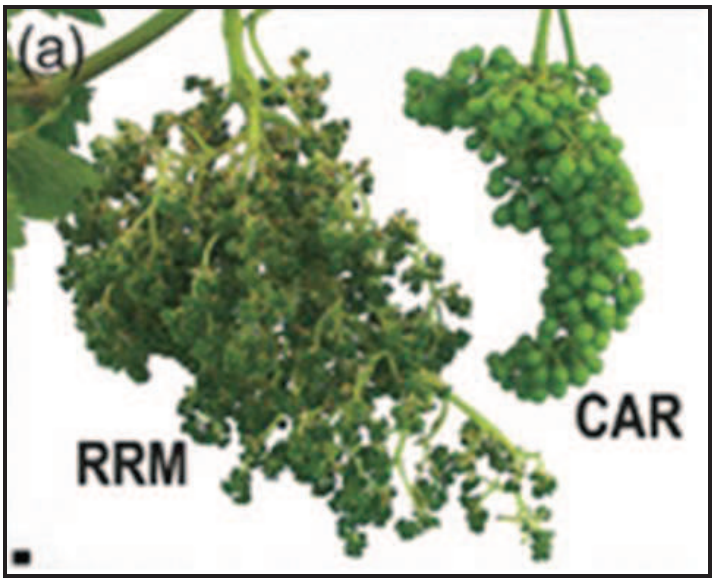


Figure 23 : Différences phénotypiques entre deux clones de Carignan : RPM, phénotype hyper ramifié et CAR, phénotype sauvage (Fernandez *et al.,* 2010).

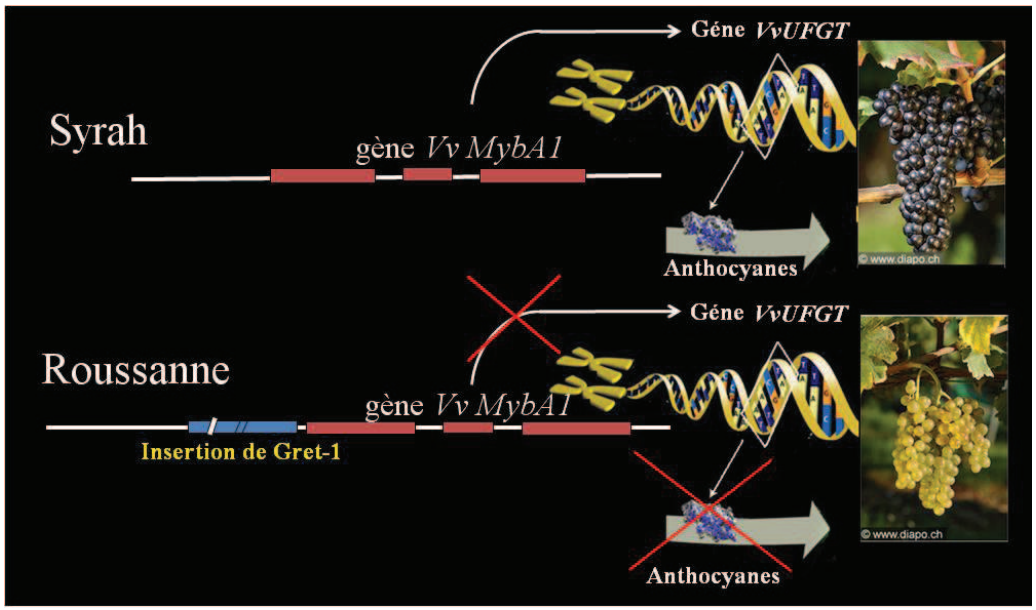


Figure 24 : Conséquence de l'insertion de l'élément *Gret-1* dans la région promotrice du gène *VvMybA1*.

Les événements de transposition des éléments transposables sont régulés par différentes composantes, en particulier les ARN interférents (Slotkin *et al.*, 2009) ou les méthylations (Slotkin & Martienssen, 2007). Les ARN interférents sont capables de complémenter les ARN spécifiques des éléments transposables empêchant la traduction de ceux-ci et favorisant leur dégradation. La méthylation de l'ADN favorise la contraction de la chromatine empêchant ainsi aux gènes des éléments transposables d'être transcrits. Ces mécanismes de régulation de la transposition varient selon l'environnement. En état de stress (biotique, abiotique), l'activité des éléments mobiles est augmentée. Par exemple chez la gueule de loup *(Antirrhinum majus)*, il a été montré que l'élément *Tam-3* se transpose mille fois plus souvent à 15 °C qu'à 25°C (Hashida *et al.*, 2003). La multiplication végétative par culture *in vitro* est une méthode connue pour entrainer un stress important et une sur-activité des éléments mobiles. Chez le riz, les plantes régénérées par culture *in vitro* comportent entre 5 et 30 nouvelles insertions de *Tos17* (Hirochika *et al.*, 1996). Chez la vigne, il a été montré par Moisy (2008) que, en plus de la culture *in vitro*, les pratiques culturales telles que la taille favorisent la transposition de certains éléments transposables.

Impact de l'activité des éléments mobiles dans les génomes

L'activité des éléments mobiles a de multiples impacts sur le génome de l'individu. Bien qu'ils aient été longtemps considérés comme faisant partie de « l'ADN poubelle », ils jouent en fait de nombreux rôles dans la structure et la plasticité du génome. Ils sont les constituants des centromères (Jiang *et al.*, 2003) et participent au maintien des télomères (Pardue *et al.*, 2005). Ils sont un des moteurs de l'expansion de la taille des génomes. Kidwell (2002) a montré que la taille des génomes de différentes espèces était corrélée à la quantité d'éléments mobiles détectés (Figure 21). Par exemple, on a constaté le doublement de la taille du génome d'*Oryza australiensis* au cours des 3 millions d'années qui viennent de s'écouler. Cet accroissement a été provoqué essentiellement par l'insertion de nouvelles copies d'éléments transposables (Piegu *et al.*, 2006). L'expansion rapide d'un élément mobile dans un génome peut être expliquée par un transfert horizontal. Quand un élément s'insère pour la première fois dans un génome ou qu'un important réarrangement survient, cet élément peut se multiplier exponentiellement avant que les mécanismes de régulation aient eu le temps de se mettre en place et neutralisent sa transposition (Figure 22). De nombreux exemples de transfert horizontaux ont été montrés chez les animaux, quelques-uns chez les plantes comme par exemple *Rider* chez la tomate (Cheng *et al.*, 2009).

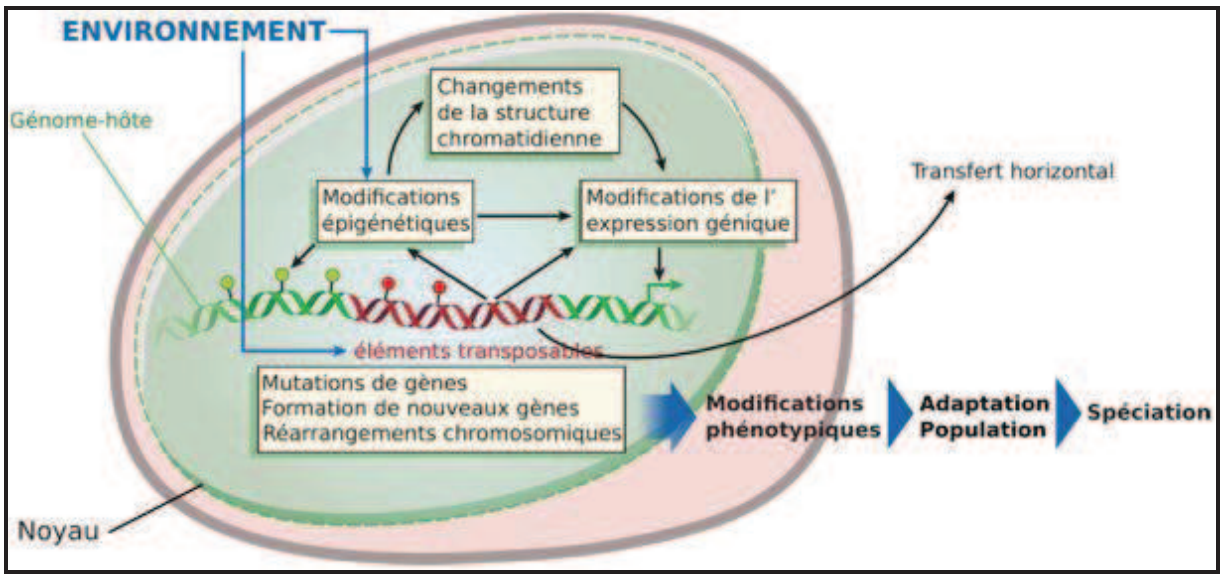


Figure 25 : Vue d'ensemble de la régulation et des impacts des éléments transposables (Moisy, 2008).

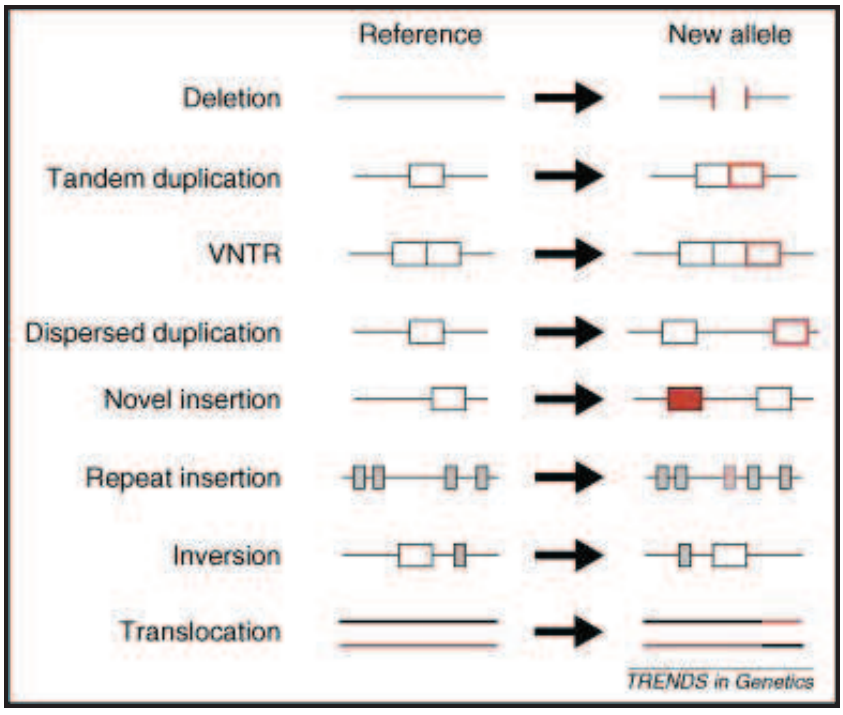


Figure 26 : Exemples de variations structurales (Hurles *et al.,* 2008).

Les insertions d'éléments mobiles ou leurs éliminations dans un gène, ou en amont de celui-ci, peuvent générer des mutations ayant un effet direct sur le phénotype. La plupart de ces mutations sont délétères au bon fonctionnement de la protéine, mais certaines peuvent au contraire engendrer de nouvelles fonctions (Long *et al.*, 2003). Chez le riz, comme chez la vigne, environ 12% des gènes comportent des éléments mobiles (Jaillon *et al.*, 2007; Sakai *et al.*, 2007). Ils contribuent de manière importante à la régulation directe de l'expression de nombreux gènes. Dans le génome humain, 3% des gènes ont leur régulation influencée par les éléments mobiles (Van de Lagemaat *et al.*, 2003). Chez la vigne, le retrotransposon *Hatvine1-rrm* entraine la surexpression du gène *VvTFL1A* et provoque une sur-ramification de la grappe (Figure 23 ; Fernandez *et al.*, 2010). Le retrotranposon *Gret-1* inséré dans le promoteur inhibe par contre l'expression du gène *VvMybA1* entrainant la non pigmentation des baies (Figure 24 ; Kobayashi *et al.*, 2004). De par leur capacité à générer des mutations et la régulation fine de leur transposition, les éléments mobiles constituent un mécanisme d'adaptation très important pour les organismes (Figure 25). Leur ancienne dénomination en ADN poubelle est donc aujourd'hui bien désuète.

2.2.1.3-Les variations structurales

Les variations structurales sont des événements mutationnels de grande ampleur (supérieure à 1Kb et pouvant atteindre plusieurs Mb), (Inoue & Lupski, 2002). Depuis quelques années, grâce aux premiers résultats produits par le séquençage massif de génomes (1000 génomes humains (Kaiser, 2008), 1001 génomes d'arabidopsis (Weigel & Mott, 2009)), plusieurs larges insertions, délétions, duplications, inversions ou translocations polymorphes entre les individus ont été identifiées (Figure 26). Les inversions et les translocations d'une large séquence d'ADN peuvent modifier l'expression d'un gène en le déplaçant dans une autre région du génome (Iafrate *et al.*, 2004; Feuk *et al.*, 2006). Les duplications « copier-coller » de séquence d'ADN peuvent entrainer un polymorphisme du nombre de copies de gènes (Copy Number Variation, CNV) dans le génome pouvant avoir un impact sur le phénotype (Sebat *et al.*, 2004; Stranger *et al.*, 2007). L'importance de ce type de polymorphisme ne cesse de croitre et certaines de ces variations structurales sont associées à différentes maladies chez l'Homme comme par exemple des maladies mentales (Sharp *et al.*, 2007) ou le diabète (Mefford *et al.*, 2007).

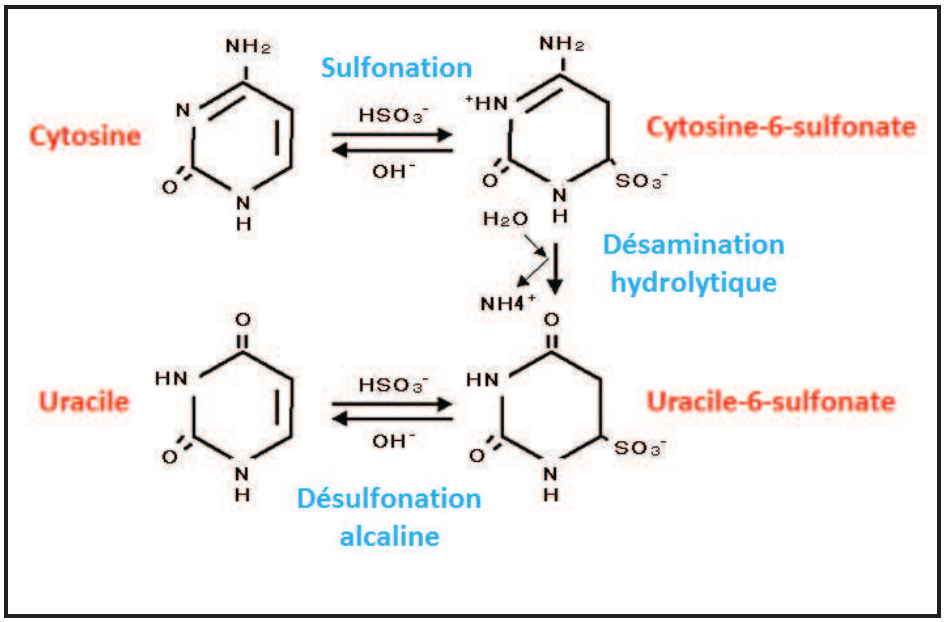


Figure 27 : Utilisation du bisulfite pour identifier les nucléotides méthylés.

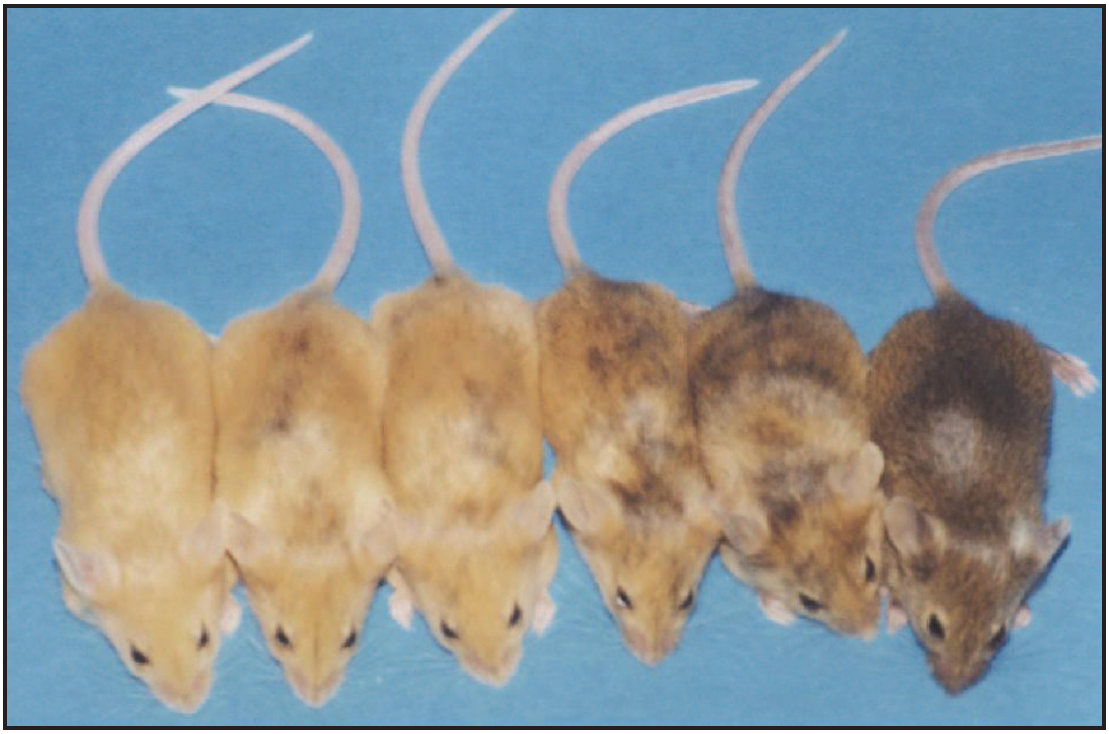


Figure 28 : Conséquences de la variation de méthylation de l'ADN dans une région régulatrice du gène *agouti* chez la souris (Morgan *et al*., 1999).

2.2.2-Les mutations épigénétiques

Les mutations épigénétiques, au contraire des mutations génétiques, n'affectent pas directement la séquence de l'ADN mais sa structure et son organisation ; on parle ainsi d'épi-génome (Holliday, 2006). L'épi-génome est un niveau de régulation récemment découvert dont on commence à peine à comprendre les différents mécanismes et leurs impacts sur le phénotype. Les modifications d'histones, les variations de méthylations de l'ADN et les petits ARN (siRNA, miRNA) peuvent entrainer des modifications épigénétiques (Rapp & Wendel, 2005). Actuellement, les variations de méthylations affectant le phénotype sont les phénomènes épigénétiques les plus étudiés (Holliday, 2006).

La méthylation de l'ADN correspond à l'ajout d'une fonction méthyle au nucléotide cytosine. L'ajout de cette fonction peut entrainer une compaction de l'ADN, rendant ainsi inaccessible aux enzymes les sites de transcription. Des enzymes responsables de la méthylation - démétyhlation de l'ADN ont été découverts chez les plantes (Agius *et al.*, 2006; Morales-Ruiz *et al.*, 2006). Ces protéines sont actives après des stimuli précis et leur activité est tissu spécifique.

La mise en évidence des profils de méthylation d'un tissu (épi-genome) est assez récente. Différentes méthodes permettent de le mettre en évidence : utilisation d'enzymes de restriction sensibles et insensibles à la méthylation ; utilisation de protéines ayant une affinité différente selon si l'ADN est méthylé ; utilisation de molécules chimiques comme le bisulfite qui convertit le nucléotide méthylé (Cytosine) en un autre nucléotide (Uracile) (Figure 27) (Piperi & Papavassiliou, 2011). C'est cette dernière méthode qui est actuellement la plus utilisée (Kankel *et al.*, 2003).

La méthylation de l'ADN joue un grand rôle dans sa compaction. Selon l'état de la chromatine (euchromatine non-compactée – hétérochromatine compactée) l'expression des gènes, des éléments transposables, et des petits ARNs varie ce qui engendre des modifications parfois importantes du phénotype (He *et al*., 2011). Par exemple, le phénotype « agouti » chez la souris est causé par un différentiel de méthylation dans la zones régulatrices du gène *agouti* (Figure 28 ; Morgan *et al.*, 1999). Finalement, la régulation des méthylations est très complexe, et au sein d'un même individu, les épigénomes varient durant le développement de l'organisme selon les tissus et l'environnement (Bird, 2002).

3-Présentation de la thèse

3.1-Problématique

La multiplication végétative est d'une très grande importance agronomique. De nombreuses plantes telles que la pomme de terre, le manioc, la canne à sucre, la banane, l'oranger, le citronnier, le café (McKey *et al.*, 2009) possèdent une certaine diversité clonale exploitée par l'Homme. La vigne fait partie des plantes où la diversité clonale est fortement exploitée dans les programmes de sélection. Elle a permis depuis 45 ans d'obtenir des gains très significatifs pour la filière viti-vinicole. L'hypothèse la plus vraisemblable expliquant l'origine biologique de cette diversité clonale est l'accumulation de mutations somatiques. Cependant aucune étude ne recense le type et la quantité de mutations présentes entre clones chez la vigne. L'objectif de cette thèse est de dresser un panorama le plus exhaustif possible sur le type et la quantité des mutations à l'origine de la différentiation génomique présente entre clones. Cet inventaire permettra par la suite de choisir les marqueurs moléculaires les plus pertinents afin d'identifier génétiquement les clones et de fournir une liste de gènes candidats susceptibles d'être associés à des caractères d'intérêt pour la sélection assistée par marqueurs.

3.2-Stratégie d'ensemble

La vigne a été la quatrième plante dont le génome a été séquencé en totalité (Jaillon *et al.* 2007). Cette avancée a permis d'envisager des approches méthodologiques impossibles jusqu'alors, et notamement de rechercher l'origine moléculaire de la variation clonale. De plus, l'avènement des séquenceurs de deuxième génération permet d'explorer le génome avec une exhaustivité grandissante pour des coûts réduits. Notre méthodologie consiste dans un premier temps, à effectuer le séquençage le plus exhaustif possible du génome de plusieurs clones à l'aide des séquenceurs nouvelle génération. Le génome de référence de la vigne a ensuite été utilisé comme matrice afin de faciliter la reconstruction des génomes des différents clones. La comparaison de leurs séquences permettra de dresser l'inventaire des mutations génétiques en les différenciant. Dans un premier temps, la diversité clonale a été étudiée dans le cépage de référence de la vigne: le Pinot. Dans un second temps, nous avons élargi notre

étude à quatre autres cépages. Les événements mutationnels les plus pertinents ont été étudiés plus en détail sur un plus grand nombre de cépages et de clones pour confirmer leur pertinence et mieux comprendre leur dynamique, afin d'évaluer la possibilité de développer une méthodologie d'identification clonale. Seules les mutations génétiques ont été étudiées dans le cadre de cette thèse. Les mutations épigénétiques étant encore à l'heure actuelle, plus complexes à étudier du fait notamment de leurs variations selon les tissus et l'environnement.

3.3-Plan de la thèse

La première partie présente les nouvelles générations de séquenceurs ainsi que les méthodes d'analyses des données produites. Pour utiliser les séquenceurs de nouvelle génération, nous avons dû mettre en place un protocole spécifique d'extraction d'ADN nucléaire qui sera décrit dans ce chapitre.

La partie suivante présentera les résultats obtenus avec la technologie 454 GS-FLX Titanium sur l'identification des événements mutationnels présents entre plusieurs clones du cépage de référence : le Pinot. L'étude des éléments mutationnels les plus pertinents sera élargie à un échantillon plus implorant et diversifié de clones de Pinot.

Les différents types de polymorphismes observés dans la partie précédente seront étudiés avec la technologie HiSeq 2000 v2 chez d'autres clones et d'autres cépages afin de confirmer les résultats obtenus chez le Pinot. Dans une perspective d'ouverture, ces polymorphismes seront également étudiés à différents niveaux de diversité : au sein d'un même individu et entre différentes accessions du genre *Vitis*.

Finalement, la dernière partie présentera une discussion générale et les perspectives de ce travail.

Chapitre 2, Méthodes de séquençage et d'analyse

1-L'avènement des séquenceurs nouvelle génération

L'idée de développer des séquenceurs haut débit est née durant l'été 1997, portée par les avancées du séquençage humain qui étaient en cours à cette époque. Alexander Todd, James Watson, Francis Crick, and Fred Sanger vont émettre l'idée d'un séquençage parallèle massif de courtes séquences d'ADN. Dix ans plus tard, les premiers séquenceurs haut débit sont commercialisés, ouvrant de nouvelles perspectives d'exploration des génomes. Les séquenceurs de nouvelle génération vont très bientôt rendre l'étape de génotypage non limitante dans les études des organismes. L'avancée rapide de ces technologies est telle que l'on parle déjà de la 3ème génération qui devrait voir le jour dans la prochaine décennie et permettra de séquencer le génome humain en quelques heures pour moins de 1000 euros (Mardis, 2008). Cette course à la séquence est dopée par des enjeux économiques importants en particulier dans le domaine médical (Ansorge, 2009).

1.1-Les technologies actuelles (2009-2011)

Actuellement deux acteurs dominent le marché, Illumina et Roche, fournissant chacun des technologies différentes. Les premières technologies telles que Polonator ou Hélicos ont eu très peu de succès à cause du prix des machines et du temps du run de séquencage (Metzker, 2010), de ce fait nous avons choisi de ne pas les détailler dans cette thèse. La technologie Solid commercialisée en 2007, connut elle aussi peu de succès et fut peu utilisée principalement parce que les séquences produites étaient de très faible longueur (35 bp). Seules les technologies Illumina et 454 seront présentées dans cette thèse. Toutes les technologies NGS comportent une étape de préparation de l'ADN après extraction (construction de la librairie) et une étape de séquençage proprement dite (Linnarsson, 2010).

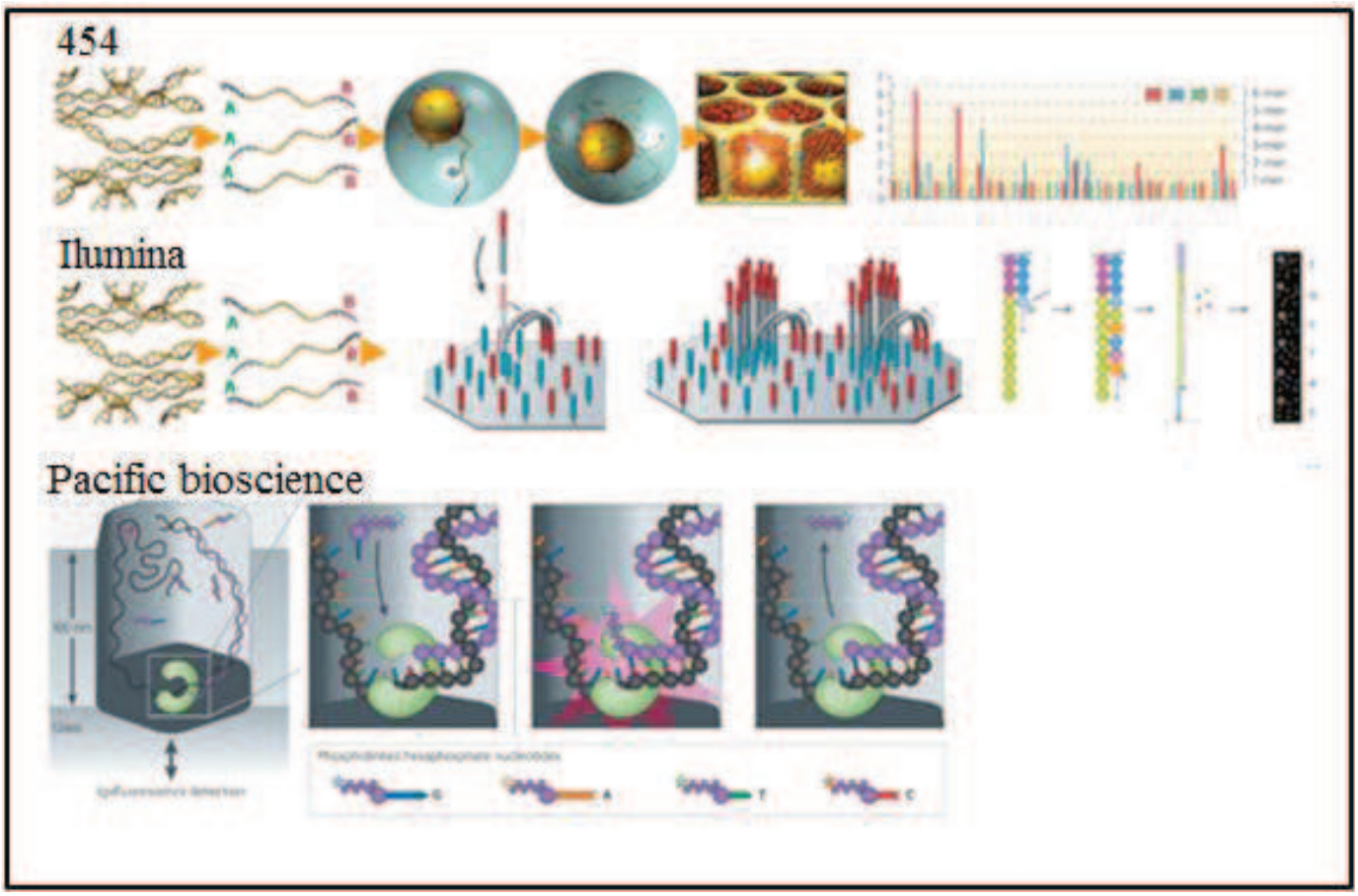


Figure 29 : Vue d'ensemble des séquenceurs 2ème et 3ème génération présentés dans cette thèse.

1.1.1-Le 454 de Roche

La technique du pyroséquencage a été présentée par Ronaghi *et al.* (1996). Le protocole et la méthode furent adaptés au haut débit et présentés par Margulies *et al.* (2005), et les premiers 454 Life Sciences GS furent commercialisés en 2005 (Figure 29).

Préparation de la librairie : L'ADN extrait est fragmenté en séquences d'environ 400 bases (Version Titanium) de façon aléatoire par nébulisation (cassure physique de l'ADN par haute pression) ou par ultra-sons. Des adaptateurs A et B sont fixés à chaque séquence et seules les séquences comportant l'adaptateur A et B (non AA et BB) se fixent sur des billes. En théorie une bille ne porte qu'une seule séquence.

Préparation du séquençage : Les billes sont ensuite placées dans une substance lipidique où aura lieu la PCR en émulsion. La PCR en émulsion consiste à enrichir la bille en séquences d'ADN à partir de l'unique molécule de départ. Cela permettra lors du séquençage d'avoir un signal de fluorescence suffisant pour être détecté par la caméra. Pour ce faire, chaque bille est entourée d'une micelle lipidique l'individualisant. La réaction PCR se produit ainsi dans chaque micelle indépendamment.

Séquençage : Les billes enrichies de molécules d'ADN simple brin sont déposées sur une plaque micro-perforée. En théorie, chaque puits ne contient qu'une seule bille, le nombre de puits étant de deux millions (Version Titanium). Un cycle de séquençage se déroule de la façon suivante :

1- Injection d'un nucléotide A marqué
2- Complémentation du/des nucléotides A
3- Elimination des nucléotides non complémentés
4- Excitation des nucléotides par un laser et réception de la fluorescence
5- et ainsi de suite pour les 3 autres nucléotides

Il se produit 200 cycles durant le séquençage (Version Titanium). En moyenne, le fragment d'ADN séquencé mesure 400 bases plus ou moins 200. En effet lors de la complémentation des nucléotides, si 3 A se suivent, alors les 3 A vont être séquencés lors d'un seul cycle de séquençage. Lorsque 3 molécules du même nucléotide sont omplémentées, le signal lumineux est en théorie 3 fois plus fort, ce qui indique que la séquence comporte 3 A

à ce niveau. Cela permet de séquencer des fragments d'ADN plus longs que le nombre de cycles de séquençage. Cependant pour un nombre de répétitions importants d'un même nucléotide (homopolymère), le signal lumineux devient difficile à interpréter et représente alors une source d'erreur importante.

1.1.2-Les séquenceurs d'Illumina

En 1997, deux chimistes anglais, Balasubramanian et Klenerman, mettent au point une approche de séquençage d'une molécule d'ADN fixée à des microsphères. Trois ans plus tard, le premier prototype haut débit, basé sur la formation de colonie d'ADN sur une puce de silice, est construit. En 2006 Illumina commercialise ses premiers séquenceurs de nouvelle génération (Figure 29).

Préparation de la librairie : Comme pour la méthode 454, l'ADN extrait est fragmenté en séquences de 400 bases (HiSeq) de façon aléatoire par nébulisation ou ultra-sons. Une sélection fine des fragments d'une taille de 400 bases (plus ou moins 100) est effectuée sur gel d'agarose. Des adaptateurs A et B vont être ensuite fixés à chaque séquence.

Préparation du séquençage : les molécules d'ADN simple brin sont fixées à l'aide des adaptateurs sur une plaque de silice comportant les adaptateurs complémentaires (Figure 29). Les séquences sont ensuite amplifiées par PCR afin de former des colonies ou des clusters sur la plaque de silice. Lors du séquençage, chaque séquence composant la colonie va émettre de la fluorescence permettant au signal d'être suffisamment fort pour qu'il soit mesurable par les détecteurs.

Séquençage : Les colonies ainsi formées sont tout d'abord dénaturées pour n'avoir qu'un seul brin d'ADN. Les cycles de séquençage se déroulent de la façon suivante :

1- Injection d'un nucléotide fluorescent
2- Complémentation d'un seul nucléotide
3- Elimination des nucléotides excédentaires
4- Excitation des nucléotides complémentés par un laser et mesure de la fluorescence
5- De même pour les 3 autres nucléotides

La taille du fragment séquencé est ici égale au nombre de cycles de séquençage. Ainsi pour 100 cycles, le fragment séquencé fera exactement 100 bases. Le problème des

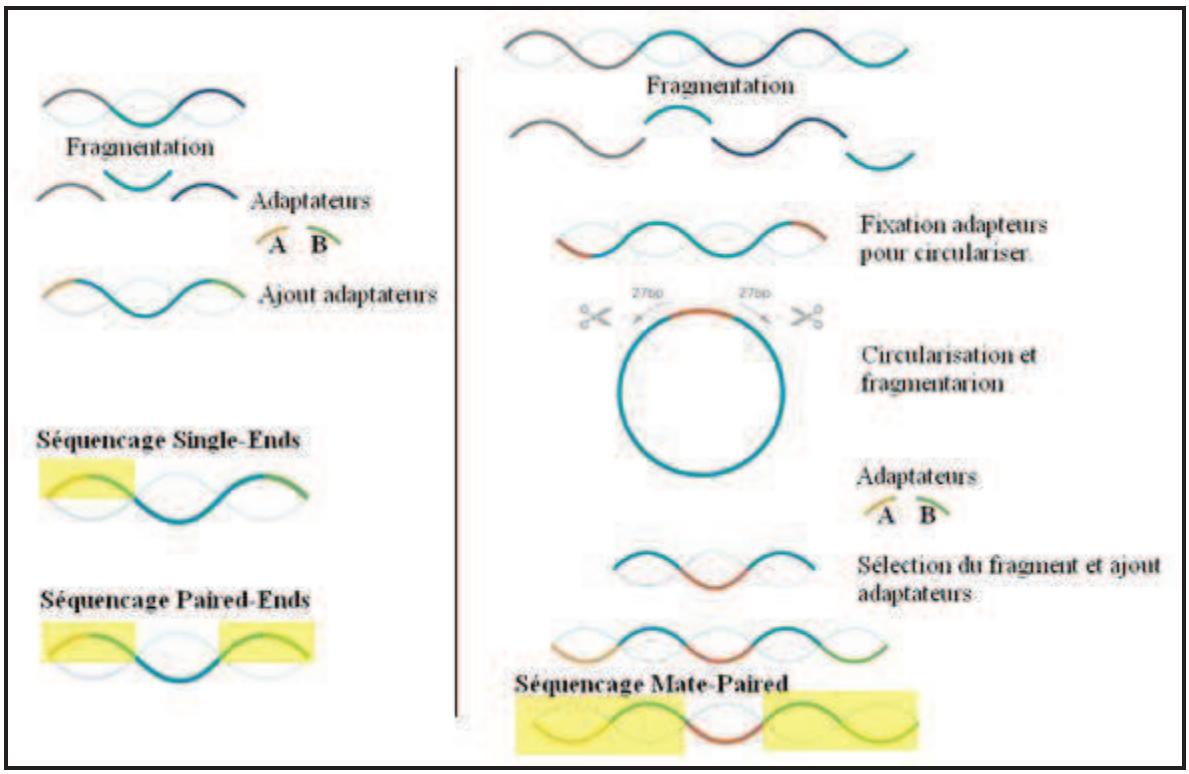


Figure 30 : Différentes options de préparation de librairies et de séquençage.

homopolymères du 454 n'est donc pas possible mais pour un même nombre de cycle, les séquences sont plus courtes.

1.2-Options de séquençage

Ces séquenceurs sont capables de séquencer de façon générale tous types de molécules d'ADN. Selon les applications et les analyses que l'on souhaite effectuer, les protocoles d'extraction d'ADN et de la fabrication de la librairie sont légèrement différents. Certains d'entre eux demandent des précautions particulières et des étapes de préparation supplémentaires. Des protocoles particuliers ont été établis pour la construction de librairie pour de l'ADN ancien (Green *et al.*, 2009), un mélange d'individus (Margraf *et al.*, 2010), de l'ARN (Meyer *et al.*, 2009) ou pour l'épi-génome (Taylor *et al.*, 2007). Les applications, une fois les données obtenues, sont immenses : étude de la diversité, recherche de mutations causales, génétique d'association, génomique comparative, épigénétique (Nordborg & Weigel, 2008; Bräutigam & Gowik, 2011).

En 2005, il n'était possible que de séquencer l'extrémité d'un fragment d'ADN (« Single-End »). Petit à petit, d'autres options de séquençages sont apparues, et actuellement 3 grands types sont proposés sur les différentes machines : "Single-End", "Paired-End", "Mate-Paired" (Figure 30). Le "Paired-End" consiste au séquençage des deux extrémités d'un même fragment d'ADN, pour lequel il reste au milieu une zone non séquencée. De plus, il peut être utilisé afin de détecter plus facilement des variations structurales ou l'insertion d'éléments transposables (Korbel *et al.*, 2007). Le "Mate-Paired" permet le séquençage des extrémités de très grands fragments d'ADN tels que ceux provenant des banques BAC (Figure 30). Il est utilisé principalement pour reconstruire le génome *de novo* d'organismes pour lesquels il n'y a pas encore de séquence de référence.

1.3-Comparaison des technologies

Les deux technologies présentent des différences sur la quantité et la longueur des séquences produites. De façon générale, le 454 produit un plus petit nombre de séquences mais celles-ci sont de plus grandes tailles (Figure 31). Au contraire, l'Illumina produit des séquences de tailles plus réduites mais en plus grand nombre. Le nombre d'erreurs de séquençage tend à diminuer au fil des améliorations des protocoles. Le 454 fait assez peu d'erreur de substitution (1/1000b) mais de nombreuses erreurs de type indels (90%), en

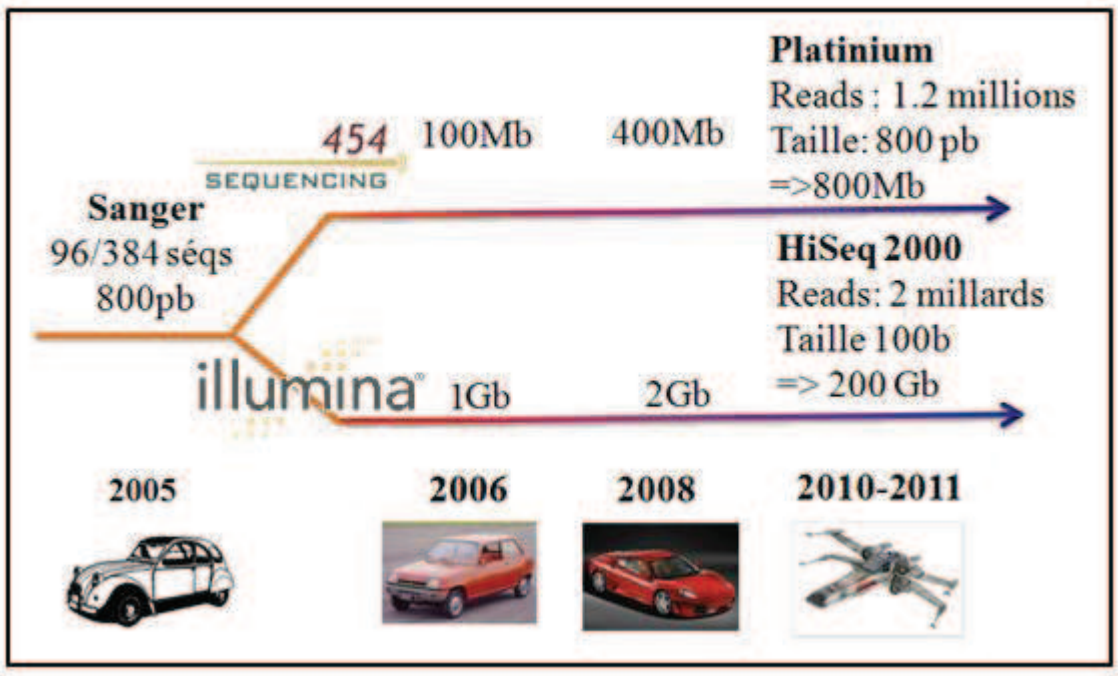


Figure 31 : Evolution des deux séquenceurs nouvelle génération les plus utilisés actuellement.

particulier générées par le problème des homopolymères (1/250b). L'Illumina au contraire ne fait pratiquement que des erreurs de substitution (95% ; 1/300b ; Harismendy *et al.* 2009 et J.M Aury, Journée Génotoul 2011). Ces taux d'erreurs changent selon la complexité de l'ADN séquencé, et des différentes versions de protocoles utilisées. La couverture permet de palier à ce taux d'erreur ainsi que le score de qualité de séquençage. Harismendy *et al.* (2009) et Nielsen *et al.* (2011) estiment que pour un fragment d'ADN séquencé dix fois (couverture 10X) avec un score de qualité de séquençage supérieur à 20 (Q20), la probabilité d'erreur est proche de zéro.

1.4-Historique des versions

L'avancée des technologies est très rapide. Environ tous les 6 mois, un nouveau kit de séquençage est disponible, ce qui permet une amélioration de la qualité du séquençage, l'augmentation du nombre de séquences, la simplification des protocoles, etc… Depuis 2005, trois grandes versions de 454 et d'Illumina ont vu le jour (Figure 31), la quantité de nucléotides séquencés a été multipliée par 8 pour le 454 et par 25 pour l'Illumina. Actuellement (Eté 2011), l'Illumina en version HiSeq 2000 permet d'obtenir au moins deux milliards de reads de 100 bases (200 Gb), et le 454 en version Platinum permet d'obtenir au moins 1,2 millions de reads de longueur moyenne de 700 bases (700 Mb). Depuis peu, l'offre des séquenceurs nouvelle génération s'est enrichie de machines intermédiaires telles que le MiSeq (Illumina) et le 454 Junior (Roche) permettant pour des coûts réduits de posséder des séquenceurs moyen débit. Dans quelques années, les séquenceurs troisième génération vont voir le jour permettant d'effectuer du séquençage sans étape d'amplification. Cela permettra de diminuer les coûts et augmentera encore le débit de séquences produites. Déjà les premiers prototypes commencent à être installés. Un Pacific Bioscience Séquenceur vient d'être installé au Genoscope (J.M Aury, Journée Génotoul 2011). Il permet d'effectuer du séquençage de fragments de 4 Kb. Le taux d'erreur reste cependant encore élevé (de l'ordre de 10%) (Figure 29).

2-Méthode d'extraction d'ADN pour les séquenceurs nouvelle génération

Toutes les molécules d'ADN fournies au NGS sont séquencées sans *a priori*. Une sélection des molécules d'ADN d'intérêt est donc nécessaire. Dans notre étude, nous nous intéressons exclusivement à l'ADN nucléaire. Nous désirons de ce fait éliminer l'ADN cytoplasmique (composé de l'ADN mitochondrial et chloroplastique) qui est présent en multi-copies dans une cellule et peut représenter jusqu'à 20% de l'ADN total (Boffey & Leech, 1982). En 2009, aucun protocole d'extraction d'ADN adapté aux contraintes des NGS n'était disponible pour les plantes. Nous avons ainsi mis en place un protocole d'extraction d'ADN répondant aux contraintes des NGS et limitant la quantité d'ADN cytoplasmique extraite. Ce protocole a été publié en janvier 2011 dans *American Journal of Botany*. Il a été utilisé pour toutes les extractions d'ADN qui ont servi pour les séquençages en NGS au cours de cette thése. Il a été également utilisé avec succès dans d'autres projets notamment pour le séquençage du génome du café. Il est utilisé comme protocole de référence sur la plateforme GénoToul (INRA Toulouse) et au Genome Analysis Center (Norwich-UK).

Nous présentons en supplément de l'article, la figure 32 qui résume les différentes étapes du protocole. Celui-ci est présenté en détail en annexe (Cf. Annexe-1). Nous avons également ajouté dans le tableau 3 la quantité de séquences d'origine cytoplasmique de tous les clones séquencés au cours de cette thèse (en 454 GS-FLX et en Illumina HiSeq 2000).

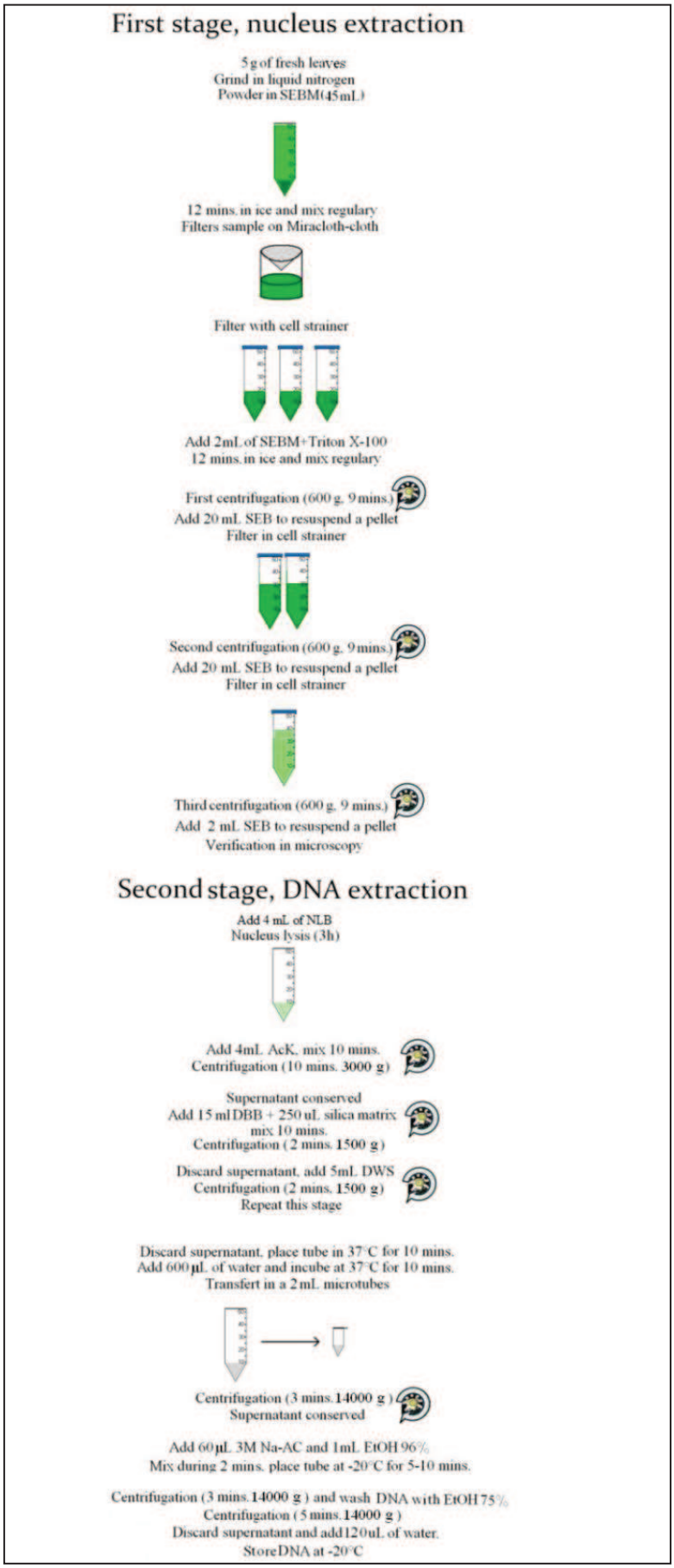


Figure 32 : Vue d'ensemble des étapes du protocole (Carrier *et al*., 2011).

American Journal of Botany

American Journal of Botany: e13–e15. 2011.

AJB Primer Notes & Protocols in the Plant Sciences

An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods[1]

Gregory Carrier[2,6], Sylvain Santoni[3], Marguerite Rodier-Goud[4], Aurélie Canaguier[5], Alexandre de Kochko[3], Christine Dubreuil-Tranchant[3], Patrice This[3], Jean-Michel Boursiquot[3], and Loïc Le Cunff[2]

[2]UMT Geno-Vigne, IFV-INRA-Montpellier Supagro, 2 place Viala, F-34060 Montpellier, France; [3]UMR 1097 DIAPC, INRA-IRD-Montpellier Supagro, 2, place Viala, F-34060, Montpellier, France; [4]UMR 1098 DAP, CIRAD-INRA-Université Montpellier II, Avenue Agropolis 34398 Montpellier France; and [5]UMR 1165 INRA-CNRS-Université d'Evry Génomique Végétale, 2 rue Gaston Crémieux 91057 Evry, France

- *Premise of the study*: In this study, we developed a nuclear DNA extraction protocol for Next Generation Sequencers (NGS).
- *Methods and Results*: We applied this extraction method to grapevines and coffee trees, which are known to contain many secondary metabolites. The nuclear DNA obtained was sequenced by the 454/GS-FLX method. We obtained excellent results, with less than 4% cytoplasmic DNA, in a similar way to a BAC (Bacterial Artificial Chromosome)–building protocol. We also compared our protocol with a classic DNA extraction using specific cytoplasmic DNA amplification. Results showed a lower cytoplasmic DNA contamination with the new protocol.
- *Conclusions*: The method presented here is fast and economical. The DNA obtained is of high quality, with a low level of cytoplasmic DNA contamination, and very efficient for the construction of sequencing libraries.

Key words: next generation sequencers; nuclear plant DNA extraction; nuclei isolation.

All protocols for Next Generation Sequencers (NGS) start with the preparation of a DNA library from extracted DNA. Quality, number, and length of the sequences produced by NGS all depend on the quality of the library that was prepared, itself dependent on the DNA quality. Generally, when using NGS, one would wish to obtain the genomic information contained in the nucleus without an excessive proportion of cytoplasmic DNA. Multiple copies of cytoplasmic genomes are present in every cell (11 to 70 copies, depending on the developmental stage and/or physiological status) (Tymms et al., 1983). The chloroplast DNA of plants, therefore, represents 17–23% of their total DNA (Boffey and Leech, 1982).

In this study, we developed a protocol that strongly reduces the amount of cytoplasmic DNA. The improvement in quality makes the method perfectly adapted to library construction for NGS. We applied this extraction method to the grapevine (*Vitis vinifera* L.) and coffee tree (*Coffea canephora* Pierre ex A. Froehner). These species are known to contain many secondary metabolites, which can generate difficulties for DNA extraction (Peterson et al., 1997; Mattivi et al., 2006).

To establish the protocol, we were partially inspired by a classic protocol for nuclear DNA plant extraction destined for the construction of BAC (Bacterial Artificial Chromosome) libraries (Peterson et al., 1997). The two stages of the protocol are: (i) isolation of nuclei and (ii) nuclear DNA extraction. Nuclear DNA obtained with this protocol was then sequenced using 454/GS-FLX technology (Roche, Basel, Switzerland) with the Titanium kit.

MATERIALS AND METHODS

All details of different steps of protocol are accessible in the supplementary methods (see Appendix S1).

For the first step, we used the "option Y" described by Peterson et al. (2000) with a small modification for nuclei isolation. The quantity of DNA necessary to build a sequencing library has to be between 2 and 5 µg, and of a quality equivalent to that necessary for BAC library construction. To obtain this DNA, we started with 5 to 6 g of leaves. Nuclei extraction is performed by first crushing the leaves quickly with a mortar and pestle in liquid nitrogen, so that enough intact nuclei are freed. A fine homogeneous powder is then obtained, which is suspended in a sucrose-based buffer containing 2-β-Mercaptoethanol and PVP to protect nuclei from oxidation. The pH is maintained relatively high to inhibit nuclease activity. The addition of Triton X-100 degrades some of the chloroplasts and mitochondria. This stage requires care, however, as it cannot be extended: lengthening the incubation would also lead to destruction of the nuclei. We used additional filtration steps compared with the protocol of Peterson et al. (2000). After filtration through Miracloth and all the centrifugations, a

[1] Manuscript received 22 September 2010; revision accepted 24 October 2010.

This work was funded by the French Ministry of Research and Higher education and the French Ministry of Food, Agriculture and Fisheries, including a grant from the IFV for GC. We are grateful to Genotoul of INRA Toulouse and Illinois University core facility for their help and advice on NGS, and to Romain Guyot and Sylvie Faure for manuscript corrections. We acknowledge Helen McCombie-Boudry for improving the English. *Authors' contributions:* GC designed the protocol, carried out the nuclei extraction, performed the grape 454 experiments and drafted the manuscript. SS designed the experiment and helped to draft the manuscript. MRG performed the microscope experiments. AC participated in the design of the protocol. AdK and CDT provided the coffee material and analyzed coffee 454 runs. JMB, PT, and LLC conceived and coordinated the study and helped to draft the manuscript. All authors read, corrected, and approved the final manuscript.

[6] Author for correspondence: gregory.carrier@supagro.inra.fr

doi:10.3732/ajb.1000371

Clones	454 GS-FLX		Illumina HiSeq 2000	
	Chloroplastique	**Mitochondrial**	**Chloroplastique**	**Mitochondrial**
777	2.3	1.4	2.5	1.1
583	1.8	1.6	3.6	2.0
386	2.1	2	3.4	2.2
E52			2.2	1.1
PNst			2.4	1.1

Table 3 : Pourcentage de séquences d'origine cytoplasmique obtenues après séquençage.

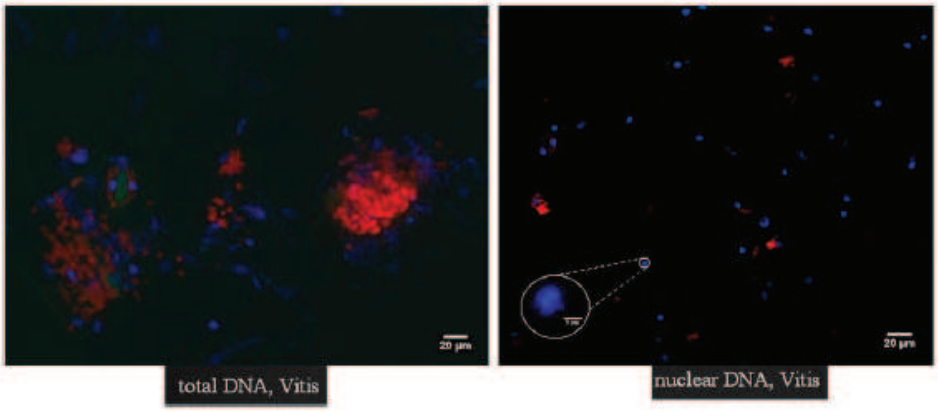


Fig. 1. Picture of nuclei grapevine extract with in DAPI coloration. Two pictures taken using a fluorescence microscope. Blue corresponds to DNA, red to chlorophyll, and green to fragment wall. Intact nuclei (blue spheres) and a few chloroplasts (red spheres) are visible on this grapevine extraction.

filtration step is done with a cell strainer (d: 40 µm). The different stages of slow centrifugation allow the nuclei to be precipitated in the bottom of the tube, whereas fragments, chloroplasts, and mitochondria remain at the surface and are then eliminated by filtration. After nuclei isolation, we examined a drop of the suspension under a fluorescence microscope with DAPI. Figure 1 shows the large numbers of intact nuclei present in this suspension. We compared the quality of visible chloroplasts in DAPI before and after nuclei purification and noted a significant reduction in chloroplast number.

The second step of the protocol was nuclear DNA extraction. The objective of preparation for a BAC library is to remove high molecular weight nuclear DNA. Nuclear DNA for NGS does not require this to be done because the DNA used for NGS is fragmented during library preparation. DNA extraction of a nuclei suspension can be made with a commercial kit such as DNeasy Plan Maxi Kit from Qiagen (Hilden, Germany). In this study, we preferred to use a protocol in which we had complete control of all the steps, to obtain DNA of high quality. In this procedure, the nuclei suspension obtained is then treated by proteinase K in the presence of lauryl sarkosyl detergent, which helps membrane lysis and the denaturation of cytoskeletal proteins, thus permitting the freeing of nuclear DNA. To purify the DNA, a precipitation with potassium acetate (salting out) eliminates broken cells and a maximum of distorted proteins. Specific absorption of DNA is realized on silica matrix (Boom et al., 1990). The final impurities are then eliminated by a succession of several washes. Purified DNA is then cleared of any alien element and is in an optimal condition for use and conservation.

The quantity of purified nuclear DNA was measured by the PicoGreen method (Murakami and McCaman, 1999). For grapevine and coffee tree, 12 and 7 µg of nuclear DNA were extracted, respectively, from 5 g of leaves of each species. Protocols for DNA library construction for the 454/Titanium (Roche) sequencing technology recommend using 2 to 5 µg of DNA. The improved protocol we present here is therefore perfectly adapted to these quantities.

We used 454/GS-FLX Titanium as the NGS methodology to test our protocol, but it is suitable for any other sequencing method. The 454 runs were successful. For grapevine two runs were made, and we obtained 988 669 and 1 052 396, sequences, with mean lengths of 350 and 360 bp, respectively. For coffee tree, a total of six runs were made. On average 1 325 441 sequences per run were obtained for a mean length of 383 bp. These excellent results, although they also depend on the quality of the sequencing libraries, are mainly due to a high standard of the purified DNA.

Cytoplasmic reference sequences are not available for coffee tree. Therefore, to quantify the level of contaminant cytoplasmic DNA sequenced, we worked on the grapevine dataset only. A BLAST search was conducted on grapevine reference sequences from the NCBI website (http://www.ncbi.nlm.nih.gov, *Vitis vinifera* chloroplast: NC 007957; *Vitis vinifera* mitochondrion: NC 007762). Out of all sequences produced in the tests of our protocol, we found that, on average, 1.7% were chloroplastic-like and 2.0% were mitochondrial-like (E-value < 1e-50 and percentage identities > 85%). These results show an elimination of a large proportion of cytoplasmic DNA by our protocol.

To compare this protocol with standard DNA extraction (Qiagen DNeasy plant), we used primer sets that specifically amplified the ATPi and RPS16 chloroplastic genes (Heinze, 2007) on four different concentrations of DNA (1 ng/µl, 0.1 ng/µl, 0.01 ng/µl and 0.001 ng/µl). Results clearly indicated no amplification from 0.1 ng/µl with our protocol, while amplification was achieved at this concentration using the classic DNA extraction method. There is no amplification from 0.01 ng/µl using the classic DNA extraction method (Fig. 2). This level of cytoplasmic DNA is similar to those obtained using BAC-building protocols (Noir et al., 2004; Zharkikh et al., 2008). Cytoplasmic DNA is still sufficiently present to be sequenced during an NGS sequencing run. This aspect could be valuable for gaining access to genomic information about cytoplasmic DNA in an NGS run, without having an excessive proportion of these sequences.

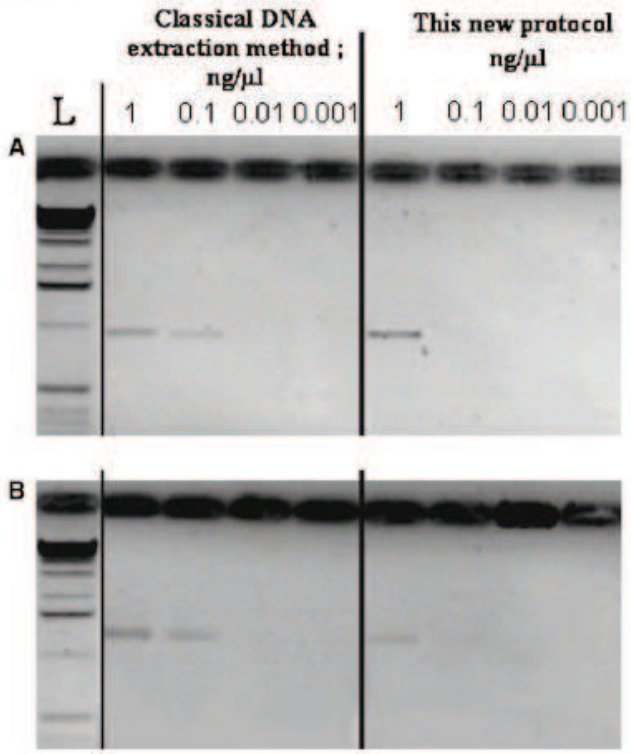


Fig. 2. Estimation of the elimination of chloroplastic DNA using the new method compared with a classic DNA extraction protocol. (A) Specific amplification of ATPi. (B) Specific amplification of RPS16, from different DNA concentrations: 1ng/µl, 0.1 ng/µl, 0.01 ng/µl, 0.001 ng/µl with DNA classic extraction.

CONCLUSIONS

The protocol presented here is fast, economical (compared with a purchased kit), and does not require the use of dangerous products. It provides adequate quantities of high-quality nuclear DNA, low in cytoplasmic contaminants, and efficient for constructing sequencing libraries. This protocol can be adapted for DNA preparations from many other plants containing high levels of phenolic compounds or polysaccharides, and it is suitable for all deep-sequencing technologies.

LITERATURE CITED

Boffey, S. A., and R. M. Leech. 1982. Chloroplast DNA levels and the control of chloroplast division in light-grown wheat leaves. *Plant Physiology* 69: 1387–1391.

Boom, R., C. J. Sol, M. Salimans, C. Jansen, P. Werthem van Dillen, and J. van des Noordaa. 1990. Rapid and simple method for purification of nucleic acids. *Journal of Clinical Microbiology* 28: 495–503.

Heinze, B. 2007. A database of PCR primers for the chloroplast genomes of higher plants. *Plant Methods* 3: 4 .

Mattivi, F., R. Guzzon, U. Vrhovsek, M. Stefanini, and R. Velasco. 2006. Metabolite profiling of grape: Flavonols and anthocyanins. *Journal of Agricultural and Food Chemistry* 54: 7692–7702.

Murakami, P., and M. T. McCaman. 1999. Quantitation of adenovirus DNA and virus particles with the PicoGreen fluorescent dye. *Analytical Biochemistry* 274: 283–288.

Noir, S., S. Partheyron, M. C. Combes, P. Lashermes, and B. Chalhoub. 2004. Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). *Theoretical and Applied Genetics* 109: 225–230.

Peterson, D., K. Boehm, and S. Stack. 1997. Isolation of milligram quantities of nuclear DNA from tomato (*Lycopersicon esculentum*), a plant containing high levels of polyphenolic compounds. *Plant Molecular Biology Reporter* 15: 148–153.

Peterson, D. G., J. P. Tomkins, D. A. Frisch, R. A. Wing, and A. H. Parterson. 2000. Construction of plant bacterial artificial chromosome (BAC) libraries. *Journal of Agricultural Genomics* 5: 1–100.

Tymms, M. J., S. Scott, and J. V. Possingham. 1983. DNA content of beta vulgaris chloroplast during leaf cell expansion. *Plant Physiology* 71: 785–788.

Zharkikh, A., M. Troggio, D. Pruss, A. Cestaro, G. Eldrdge, M. Pindo, J. T. Mitchell, et al. 2008. Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: Problems and solutions. *Journal of Biotechnology* 136: 38–43.

Figure 33 : Les différents types de fichiers utilisés pour les analyses de séquences: Fasta) fichier contenant la séquence nucléotidique; Qual) fichier contenant le score qualité de séquençage du nucléotidique correspondant ; FastQ) fichier contenant à la fois la séquence nucléotidique et le score de qualité de séquençage.

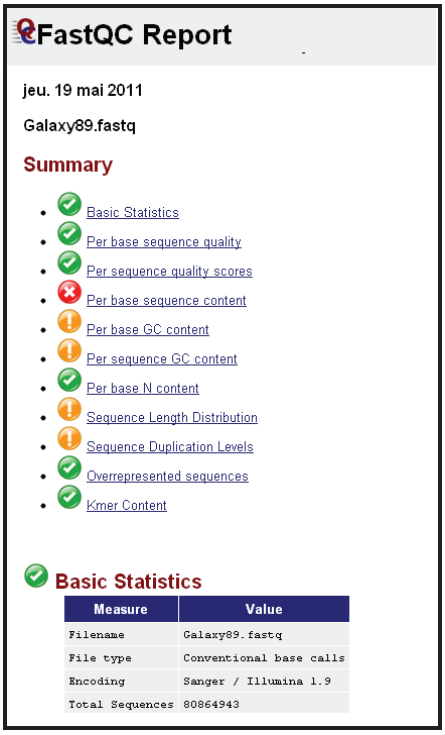
3-Etat de l'art de la bio-informatique pour les NGS

Une des difficultés des méthodes de séquençage haut débit est qu'il est désormais inimaginable d'analyser manuellement les séquences obtenues. L'analyse bioinformatique est désormais indispensable. L'avancée rapide des technologies de séquençage haut débit a eu des répercussions directes sur le développement des logiciels capables d'analyser de telles masses de données.

En 2009, au début des séquenceurs de nouvelle génération, très peu de logiciels permettant d'analyser ce type de données étaient disponibles. Aujourd'hui, compte tenu du succès des NGS, le nombre de logiciels a augmenté exponentiellement. La comparaison des différents logiciels reste assez floue et il est souvent difficile de savoir lequel utiliser de préférence. Un tri des logiciels est souvent effectué par la popularité de ces derniers. En général, les logiciels utilisés dans le programme 1000 génomes humains sont très populaires, mais ne sont pas forcément tous adaptés aux génomes des plantes. Dans ce paragraphe, nous allons rendre compte des principales étapes pour analyser les données de séquençage. Il faut souligner que la liste des logiciels cités ici correspond à ceux testés durant cette thèse et de ce fait n'est pas exhaustive (Une liste plus importante est disponible sur le site suivant : http://seqanswers.com/wiki/Software/list).

3.1- Les fichiers de sortie des séquenceurs nouvelle génération

Les données de sortie des séquenceurs sont constituées d'un fichier Fasta, comportant toutes les séquences produites et un fichier Qualité, associant pour chaque base séquencée un score de qualité. Le score de qualité de séquençage n'est pas encodé de la même façon selon les constructeurs. Ainsi le 454 est codé de 0 à 40 et l'Illumina de 64 à 120. Depuis peu, un format de fichier regroupant les deux informations a vu le jour : FastQ. Celui-ci est maintenant utilisé en routine lors des analyses NGS (Figure 33).


FastQC Report

jeu. 19 mai 2011

Galaxy89.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	Galaxy89.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	80864943

Figure 34 : Exemple d'un rapport du logiciel FastQC.

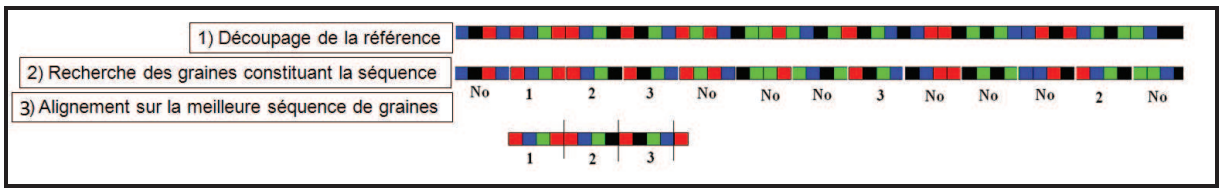


Figure 35 : Représentation simple de l'algorithme en graines (ou hash). La séquence de référence est découpée en graines. La chaîne de graines constituant la séquence donnée est recherchée parmi les graines de la séquence de référence.

3.2- Les logiciels pour traiter les fichiers et vérifier la qualité des séquences

Une des premières étapes, après avoir obtenu les données, est de vérifier la qualité du séquençage. Le logiciel FastQc (A. Simon, Babraham Institute) permet d'avoir un ensemble de statistiques permettant de connaitre la qualité du séquençage effectué (Figure 34). De façon générale, une qualité moyenne des séquences ne dépassant pas Q20 est à exclure (Nielsen *et al.*, 2011). Il est à noter que la qualité des bases diminue à l'extrémité 3' des séquences (Dohm *et al.*, 2008), ceci étant dû à la détérioration du signal lumineux en fin de séquençage. Différents logiciels disponibles permettent de filtrer ces séquences (FastX-Toolki de Hannon lab, Samtools et Picards Tools (Li, H *et al.*, 2009)) pour ne garder que les séquences de qualité que l'on estime suffisante.

3.3- Les logiciels d'alignement

L'alignement consiste à reconstruire un génome séquencé en utilisant un génome de référence comme modèle. Tous les organismes ne possèdent pas encore de génome de référence séquencé, contrairement à la vigne dont le génome de référence de qualité nous a été d'une grande utilité (http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/). A l'heure actuelle, il existe un très grand nombre de logiciels d'alignement. Ils se découpent en deux grandes classes, les logiciels basés sur un algorithme dit « de hachage » et ceux basés sur la Burow Wheler Method (Li & Durbin, 2009).

Les logiciels fonctionnant sur l'algorithme de hachage ont été les premiers à être mis en place en 2007. Deux grands types de logiciels sont disponibles, ceux qui chargent en mémoire les séquences produites et ceux qui chargent en mémoire la séquence de référence. Selon la taille du génome de référence et le nombre de séquences à aligner, l'un ou l'autre des logiciels peut être plus approprié. Le logiciel MAQ (Li H *et al.*, 2008), charge en mémoire les séquences à aligner, c'est-à-dire qu'il fragmente les séquences en petits morceaux appelés « graines ». Ces graines sont ensuite recherchées sur le génome de référence de façon à identifier la région où toute la chaine de graines se positionne. Les logiciels tel que SOAP (Li R *et al.*, 2008), BFAST (http://genome.ucla.edu/bfast/) et MosaikAssembler (http://bioinformatics.bc.edu/marthlab/) fragmentent la séquence de référence et la meilleure

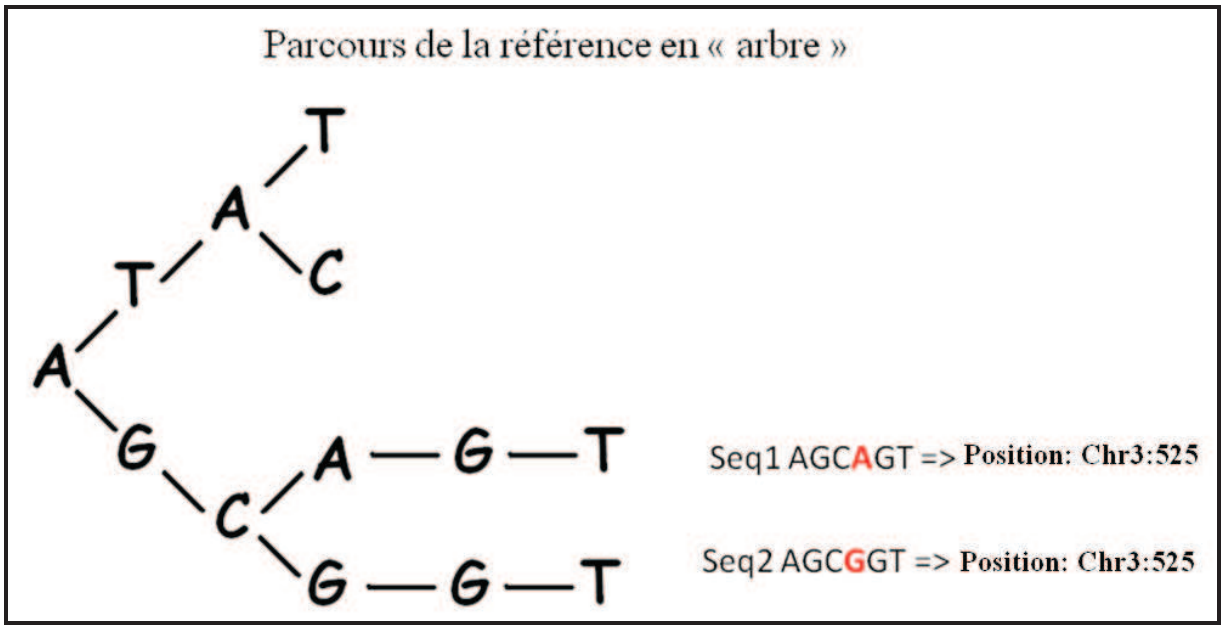


Figure 36 : Représentation simple de l'algorithme de Burrows-Wheller (Li & Durbin, 2009). Le génome de référence est parcouru sous forme d'arbre. A chaque position sur la séquence, une seule branche est choisie parmi les quatre possibles. Le nombre de branches possibles augmente proportionnellement avec le nombre d'erreurs autorisées. Cela augmente ainsi le nombre de possibilités et ralentit le temps de recherche de la séquence sur le génome de référence.

chaine de graines est recherchée pour chaque séquence (Figure 35). Actuellement au vu des débits des séquenceurs nouvelle génération, les logiciels qui chargent en mémoire la séquence de référence sont les plus utilisés.

La deuxième classe de logiciels apparue en 2009 est basée sur l'algorithme de Burrows Wheler Method (Li and Durbin 2009 ; Figure 36). Plusieurs logiciels comme BOWTIE (Langmead *et al.*, 2009), BWA (Li & Durbin, 2009) et SOAP2 (Li R *et al.*, 2009) utilisent cet algorithme. Cette méthode a l'avantage d'être extrêmement rapide pour les génomes peu polymorphes comme l'humain (1 SNP / 1042 bases (Wang *et al.*, 2008)). Pour cette raison, le logiciel BWA est rapidement devenu très populaire dans le consortium 1000 génomes humains. Cependant, pour des séquences de grandes tailles telles que celles produites en 454 ou des génomes plus complexes, les logiciels basés sur la méthode de hachage, bien que plus lents, s'avèrent plus efficaces (Flicek & Birney, 2009).

L'évolution de ces logiciels est très rapide. Actuellement la recherche se poursuit sur le développement de logiciels capables d'effectuer de l'assemblage guidé par une séquence de référence. Les régions non reconstruites par alignement simple sur la référence sont reconstruites par une méthode d'assemblage. Cela permettra, pour les régions non présentes sur la référence ou très difficiles à reconstruire par alignement, d'être accessibles.

3.4- Les logiciels de recherche de polymorphismes

La liste des logiciels permettant la recherche de polymorphismes de type SNP et indels est assez bien fournie : Gigabayes et Freebayes (http://bioinformatics.bc.edu/marthlab/), GATK (DePristo *et al.*, 2011), SamTools (Li H *et al.*, 2009). Le choix du logiciel se fera sur différentes options de filtres disponibles. Pour les individus hétérozygotes, il est important que le logiciel puisse tenir compte de la ploïdie. Il est aussi important que des filtres comme la couverture minimale, le score qualité de séquençage minimum, la fréquence minimale de l'allèle minoritaire soient paramétrables pour valider ou non un SNP (Nielsen *et al.*, 2011).

Pour la recherche de variations structurales comme les éléments mobiles, les translocations, inversions, grands indels, nombre de copies de gènes, très peu de logiciels sont disponibles. Seuls quelques outils très récents adaptés à l'humain viennent d'apparaitre

Figure 37 : Présentation de la plateforme Galaxy. (*marmadais.cirad.fr*).

(Medvedev *et al.*, 2009; Hormozdiari *et al.*, 2010). Dans un avenir proche le nombre de ce type de logiciels devrait augmenter.

3.5- Pipeline et Genome Browser

La mise en place d'un pipeline est toujours complexe. Récemment, une interface appelée Galaxy a été développée à l'occasion du projet 1000 génomes humains (Figure 37 ; Blankenberg *et al.*, 2001). Celle-ci est « open source » et permet avec une très grande facilité d'utiliser les outils installés dans l'interface et de les enchainer sous la forme d'un pipeline. Une fois tous les résultats obtenus, une validation manuelle du polymorphisme sur les gènes candidats est généralement souhaitable. Plusieurs « Génome Browser » ont vu le jour permettant de visualiser facilement les données NGS ainsi que tous les autres types de données (annotations, données d'expression etc…). Actuellement IGV (Robinson *et al.*, 2011) et Tablet (Zhang *et al.*, 2010) sont les « Génome Browser » les plus populaires.

4-Conclusion

Les nouvelles générations de séquenceurs ont permis de rendre les étapes de génotypage non limitantes et permettent d'envisager des études sur génomes entiers (Nordborg & Weigel, 2008). Les projets de re-séquencage d'envergure ont ainsi pu voir le jour (1000 génomes humains, (Kaiser, 2008) et 1001 génomes d'*Arabidopsis* (Weigel & Mott, 2009), etc…) permettant d'effectuer des comparaisons de génomes afin d'identifier du polymorphisme moléculaire lié avec certains polymorphismes phénotypiques. Les différentes préparations d'ADN, ARN ont du être adaptées à ces nouveaux séquenceurs. De nouveaux protocoles de préparation d'ADN sont mis en place permettant ainsi d'étudier par exemple les méthylations de l'ADN (Taylor *et al.*, 2007). L'arrivée de tels volumes de données nécessite ainsi le développement de nouvelles méthodes d'analyses informatiques et la mise en place d'infrastructures informatiques d'envergure. Le développement d'interface telle que Galaxy permet de faciliter l'utilisation des différents logiciels d'analyse. Au cours de ma thèse j'ai eu aussi l'occasion de suivre et de participer directement à cette révolution du génotypage apportée par ces nouvelles technologies.

Chapitre 3, La variation clonale chez le Pinot

1-Introduction

Ce chapitre est consacré à l'identification des événements mutationnels présents entre deux clones. Pour cela, plusieurs clones de Pinot ont été séquencés et comparés pour identifier et quantifier d'éventuels polymorphismes. Les éléments mutationnels les plus pertinents ont par la suite été étudiés dans un échantillon beaucoup plus large de la diversité du Pinot.

2-Etude du cépage de référence de la vigne, le Pinot

Le nom du cépage Pinot proviendrait de la forme de ses grappes en pomme de pin. C'est un cépage cultivé depuis très longtemps. Il serait apparu dans l'Est de la France durant le premier siècle après JC (Mc Govern, 2003) et a donné naissance à de nombreux cépages de grande renommée comme le Chardonnay, le Gamay, l'Aligoté…(Bowers *et al.*, 1999). Le Pinot noir a une grande importance économique, c'est le 11éme cépage planté dans le monde (J.M Boursiquot Com. Pers.). En 2006, 28 006 hectares de Pinot noir sont cultivés en France, essentiellement en Bourgogne et en Champagne, contre 11 876 hectares en 1968 (Boursiquot *et al.*, 2007; OIV, 2007). Ce cépage est à la source des grands crus de Bourgogne comme par exemple la Romanée-Conti et le Chambertin. Il offre également une très grande diversité clonale. On dénombre actuellement 40 clones agréés de Pinot noir, deux clones de Pinot blanc, 3 clones de Pinot gris et 15 clones de Meunier. Tous ces clones possèdent des caractéristiques propres. La figure 38 présente les différences entre les clones de Pinot noir (Boursiquot *et al.*, 2007).

Le Pinot noir fut choisi comme cépage de référence de la vigne du fait de son intérêt historique et économique. Deux individus furent séquencés en 2007. Pour le séquençage du génome de référence, un individu homozygote dérivant du Pinot noir (PN40024) a été choisi pour faciliter la reconstruction *de-novo* du génome (Jaillon *et al.,* 2007). En parallèle, le clone ENTAV-INRA® n°115 fut séquencé par Velasco *et al*. (2007) avec une plus faible profondeur, 6,4X au total contre 12X pour le génome de référence.

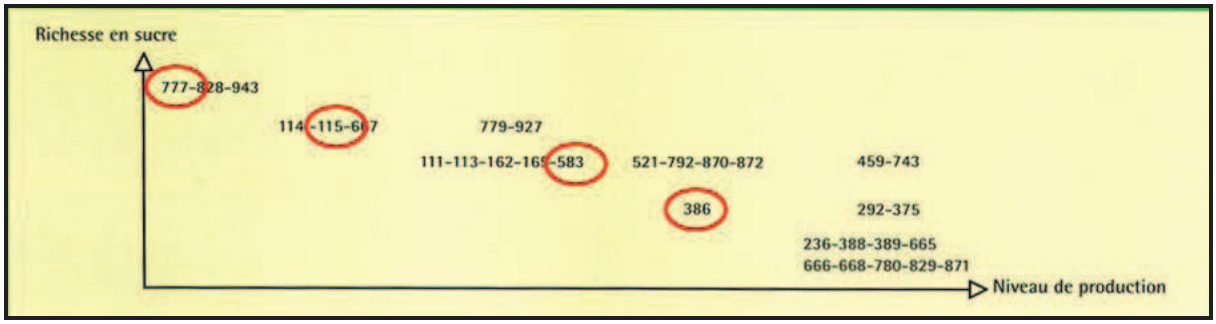


Figure 38 : Différence de niveau de production et de richesse en sucre des clones de Pinot noir agréés (Boursiquot *et al.,* 2007).

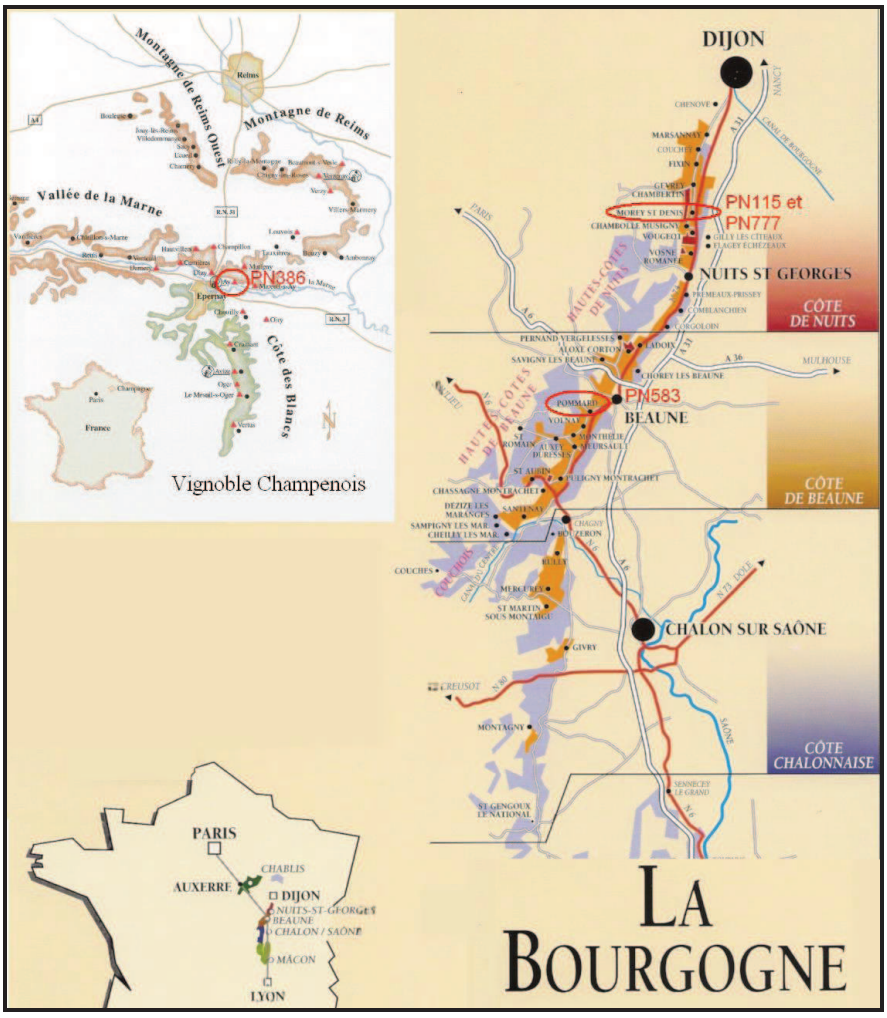


Figure 39 : Origine géographique des clones de Pinot noir agréés et séquencés.

2.1-Sélection de l'échantillon de clones

Dans notre première étude sur la diversité clonale de la vigne nous nous sommes donc focalisés sur le Pinot, car le polymorphisme structural entre le génome de référence (PN40024) et les clones devraient être relativement faibles facilitant ainsi leur reconstruction.

Pour avoir une représentation de la diversité clonale du Pinot noir, en supplément du clone n°115, trois autres clones ont été choisis pour constituer notre échantillon : Les clones ENTAV-INRA® n° 386, 583 et 777. Ceux-ci furent séquencés par la technologie 454-GS-FLX Titanium. Chacun des clones sélectionnés représente une classe phénotypique (productivité/qualité) parmi la diversité des clones de Pinot noir agréés (Figure 38). Les clones n° 115, n° 777 et n° 583 sont originaires de Côte d'ôr, plus précisément pour les deux premiers, de la commune de Morey Saint-Denis et pour le clone n° 583 de la commune de Pommard. Le clone n° 386 est d'origine Champenoise, de la commune d'Ay dans la Marne (Figure 39 ; Fiches de clones éditées par le Grapvi, 2000).

Afin de confirmer les résultats obtenus dans notre étude, une sélection de quatre éléments mutationnels a été étudiée sur l'ensemble des clones de Pinot agréés. Tous les échantillons ont été prélevés dans la collection du Domaine de l'Espiguette de l'IFV (40 clones de Pinot noir, 2 clones de Pinot blanc, 3 clones de Pinot gris et 15 clones de Meunier (Cf. Annexe-2).

3-Choix des méthodes de séquençages haut débit, 2009

En décembre 2008 lorsque ma thèse a débuté, le 454 était en version Titanium ($1,2E10^6$ séquences de longueur moyenne de 400 bases), l'Illumina était en version GaII ($30E10^6$ séquences de 50 bases). Le génome de la vigne étant de 470 Mb, un séquençage en 454 permettait d'obtenir environ 1X de couverture alors qu'un séquençage en Illumina permettait d'obtenir 3X de couverture. Malgré la différence du nombre de bases séquencées entre les deux machines, nous avons choisi d'utiliser le 454 car les séquences produites étaient de taille très supérieure à celles de l'Illumina. De plus la technologie "Paired-End" n'était pas disponible en France. Cela devait faciliter la reconstruction du génome par alignement sur le

génome de référence. En effet, le génome de la vigne est constitué de 41% d'éléments répétés (Jaillon *et al*., 2007) et est très hétérozygote (Velasco *et al*., 2007). Sur de tels génomes, pour effectuer un alignement de bonne qualité, il est recommandé d'avoir des séquences d'au moins 100 bases (Palmieri & Schlötterer, 2009). Nos séquençages 454 en version Titanium ont été effectués en collaboration avec la plateforme GénoToul à l'INRA de Toulouse. Nous avons pu effectuer nos séquençages en projet pilote et ainsi réaliser les tous premiers séquençages 454 Titanium d'Europe.

4-Identification des polymorphismes moléculaires chez le Pinot

Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera L.*

Grégory Carrier.[1], Loïc Le Cunff.[1], Alexis Dereeper.[2], Delphine Legrand. [1], François Sabot. [3], Olivier Bouchez. [4], Laurent Audeguin. [1], Jean-Michel Boursiquot.[2], Patrice This. [2]

Somatic polymorphism in grape

GC: gregory.carrier@supagro.inra.fr
[1] UMT Geno-Vigne®, IFV-INRA-Montpellier Supagro, 2 place Viala, 34060 Montpellier, France

LLC: loic.lecunff@supagro.inra.fr
[1] UMT Geno-Vigne®, IFV-INRA-Montpellier SupAgro, 2 place Viala, 34060 Montpellier, France

AD: Alexis.Dereeper@ird.fr
[2] UMR RPB, IRD-UM2-CIRAD, 911 avenue Agropolis 34394 Montpellier, France

DL: delphine.legrand@vignevin.com
[1] UMT Geno-Vigne®, IFV-INRA-Montpellier SupAgro, 2 place Viala, 34060 Montpellier, France

FS: francois.sabot@ird.fr
[3] UMR DIADE, IRD-UM2-CIRAD, 911 avenue Agropolis 34394 Montpellier, France

OB: olivier.bouchez@toulouse.inra.fr
[4] Plateforme Génomique de Toulouse Midi-Pyrénées, INRA Auzeville 31326 Castanet-Tolosan, France

LA: Laurent.Audeguin@vignevin.com
[1] UMT Geno-Vigne®, IFV-INRA-Montpellier Supagro, 2 place Viala, F-34060 Montpellier, France

JMB: boursiqu@supagro.inra.fr
[5] UMR AGAP Montpellier SupAgro, 2, place Viala, F-34060, Montpellier, France

PT: patrice.this@supagro.inra.fr
[5] UMR AGAP, INRA, 2, place Viala, F-34060, Montpellier, France

Abstract

Through multiple vegetative propagation cycles, clones accumulate mutations in somatic cells that are at the origin of clonal phenotypic diversity in grape. Clonal diversity provided clones such as Cabernet-Sauvignon N°470, Chardonnay N° 548 and Pinot noir N° 777 which all produce wines of superior quality. The economic impact of clonal selection is therefore very high: since approx. 95 % of the grapevines produced in French nurseries originate from the French clonal selection.

In this study we provide the first broad description of polymorphism in different clones of a single grapevine cultivar, Pinot noir, in the context of vegetative propagation. Genome sequencing was performed using 454 GS-FLX methodology without *a priori,* in order to identify and quantify for the first time molecular polymorphisms responsible for clonal variability in grapevine. New generation sequencing (NGS) was used to compare a large portion of the genome of three Pinot noir clones selected for their phenotypic differences. Reads obtained with NGS and the sequence of Pinot noir ENTAV-INRA® 115 sequenced by Velasco *et al*., were aligned on the PN40024 reference sequence. We then searched for molecular polymorphism between clones.

Three types of polymorphism (SNPs, Indels, mobile elements) were found but insertion polymorphism generated by mobile elements of many families displayed the highest mutational event with respect to clonal variation. Mobile elements inducing insertion polymorphism in the genome of Pinot noir were identified and classified and a list is presented in this study as potential markers for the study of clonal variation. Among these, the dynamic of four mobile elements with a high polymorphism level were analyzed and insertion polymorphism was confirmed in all the Pinot clones registered in France.

Introduction

Genomes were thought to be stable constituents of living organisms until Barbara McClintock's discovery of genome plasticity opened up a new avenue of research (McClintock, 1984). Dynamics of genomes have thus become an important field of research, SNPs and short indels being the most widely studied polymorphisms. These have a potential impact on phenotypic variations (McCarroll et al., 2008), in particular non-synonymous SNPs located in regulatory regions (McNally et al., 2006; Ramensky et al., 2002). Similarly, mobile elements drive genome evolution (Kazazian, 2004), playing an important role in mutations responsible for genomic reorganizations (Kidwell, 2002) and genome size variations (Piegu et al., 2006). In this way, 82% of the maize genome is composed of overlapping mobile elements (Schnable et al., 2009). Other mechanisms of genome regulation such as epigenetic variations (Doerfler et al., 2006; Zilberman et al., 2007) chromosome rearrangements (Eichler and Sankoff, 2003) and copy number variations (Freeman et al., 2006; Stranger et al., 2007) could also have an impact on phenotypic variations.

A significant number of domesticated plants including banana, potato, grape, coffee tree are vegetatively propagated to maintain agronomically valuable genotypes (McKey et al., 2009). However, after many propagation cycles, clones accumulate phenotypic differences in agronomic traits and clonal diversity appears (Orive, 2001). This diversity can then be used to select the best clones within a given variety. Indeed, several clonal selection programs for grape, potato or banana have led to the release of new certified clones with very significant gains for the industry. In particular, clonal diversity in grape is used to select the best clones for commercial purpose as it is the only solution to access a plant diversity without modifying the identity of cultivars with worldwide repute. Cultivar identity is crucial in the case of appellation wines in Europe which are produced from a restricted list of specific cultivars.

Vegetative propagation has been used since the end of Antiquity period (Mc Govern, 2003) and allows grape to display a remarkable clonal diversity (Schön et al., 2009). Previous studies of grapevine clonal diversity using SSR markers enabled the identification of limited clonal polymorphism in a few groups of clones (Hocquigny et al., 2004; Moncada et al., 2006). However SSR analyses are not an efficient way to distinguish genetic differences between clones (Imazio et al., 2002; Pelsy et al., 2010). Alternatively, the S-SAP approach using universal retrotransposon based primers revealed polymorphism between five Pinot clones

		PN386	PN583	PN777	Mean of 3 clones	PN115
% of aligned sequences	Alignment Step 1	48.1	42.5	40.2	43.6	57.7
	Alignment Step 2	0.9	1.1	1.0	1.0	1.3
	Alignment Step 3	12.5	21.7	16.7	16.9	8.0
	Total of aligned reads	61.5	65.3	57.9	61.5	67.0
% of unaligned sequences	Repeat elements	12.5	13.4	13.90	13.2	12.7
	Paralogs	12.0	10.1	13.0	11.7	20.3
	Cytoplasmic DNA	4.2	3.6	3.7	3.8	
	Unknown	8.1	6.8	10.4	8.4	
	Contamination (other organisms)	0.01	0.01	0.01	0.01	
	Low quality reads	1.7	0.8	1.4	1.3	
	Total of unaligned reads	38.5	34.8	42.1	38.4	33.0

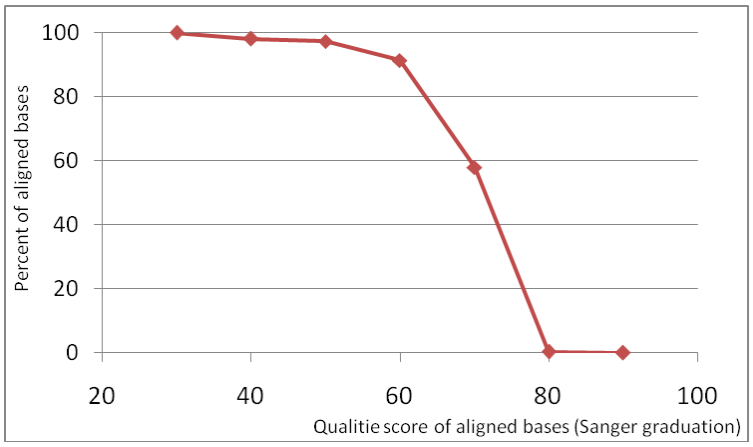
Table 1: Description of alignments on PN40024: Proportion of aligned sequences after first, second and third steps of the alignment process for the different clones and for PN115 sequences and of unaligned reads on the reference genome with their possible composition.

	PN386	PN583	PN777	Common regions	Reference genome covered	PN115 genome covered
Coverage 1X or more	113 Mb	132 Mb	139 Mb	95 Mb	194 Mb	168 Mb
Coverage 2X or more	46 Mb	64 Mb	44 Mb	16 Mb	122 Mb	98 Mb
Coverage 3X or more	15 Mb	25 Mb	14 Mb	0.2 Mb	54 Mb	52 Mb
Coverage 4X or more	6 Mb	11Mb	5 Mb	0 Mb	22 Mb	22 Mb
Coverage 5X or more	3 Mb	5 Mb	2 Mb	0 Mb	10 Mb	10 Mb
Coverage 6X or more	1.3 Mb	2.2 Mb	1 Mb	0 Mb	4.5 Mb	4.5 Mb

Table 2: Coverage of sequenced genomes: Size of the portion of genome aligned on the reference genome at different coverage levels for the three clones. Common regions correspond between all clones sequenced in 454 GS-FLX. In the polymorphism call we only considered regions with 6X coverage or more.

(Wegscheider et al., 2009) although use of *Vine-1* based primers (Verries et al., 2000) failed to reveal any variation between six Pinot clones (Labra et al., 2004). Pinot is one of the oldest grape cultivars (Boursiquot et al., 2007; This et al., 2006) and among the noblest, being used notably in Champagne and Bourgogne wines. It displays extensive clonal diversity and, in France alone, 64 different Pinot clones are certified and marketed (Boursiquot et al., 2007). Furthermore Pinot noir was the cultivar choosen in grapevine genome sequencing projects: the grape reference genome using a near homozygous line PN4024 (Jaillon et al., 2007) derived from Pinot Noir cultivar by successive selfings and the second sequencing project using Pinot noir clone ENTAV-INRA® 115 (PN115) (Velasco et al., 2007). Pinot studies can now fully benefit from existing genomic tools since the release of the reference genome sequences (Jaillon et al., 2007; Velasco et al., 2007) available through the grape genome browser (http://www.genoscope.cns.fr/)

New generation sequencing (NGS) has changed the landscape of genetics and genomics studies and allowed questions to be answered at genome scale (Mardis, 2008; Nordborg and Weigel, 2008). Until now, no study has proposed a broad description of polymorphism linked to vegetative propagation. In the present study, we thus exploited the power of NGS and the grape genomic tools to perform a genome-wide comparison of grape clone genomes without *a priori* knowledge. In order to quantify the different types of polymorphisms (SNP, indel, mobile elements) likely involved in clonal diversity, we sequenced 3 Pinot noir clones (PN386, PN583, PN777) selected for their phenotypic differences using 454 GS FLX methodology. We compared a portion of these Pinot noir clones with the available sequences of PN115 (Velasco et al., 2007) after alignment on the PN40024 reference genome. Consequences of theses polymorphisms will be discussed as well as potential uses of these results for the wine industry.



Supplementary Figure 1: Percentage of aligned bases with different quality alignment scores. 90% of aligned bases had a quality score of more than 60.

	PN386	**PN583**	**PN777**	**PN40024**
% GC in alignedsequences	36.0	35.0	35.0	33.0
% CpG in alignedsequences	2.4	2.4	2.4	2.2
% CnG in alignedsequences	0.9	0.9	0.9	0.9
% Exons	9.9	10.6	7.9	6.9

Table 3: Composition of 454 reads aligned with the reference genome. We compared the percentage of GC, CpG, CnG and exons in the 454 data set and the reference genome. Percentage of GC, CpG, CnG were estimated with a Perl script. Percentage of exons was estimated by Blast 2.0 (id > 85%) with the annotation of the reference genome dated on March 19 2010.

Results

Alignment and representation of the Pinot noir clone sequences on the reference genome

Genome reconstruction by alignment

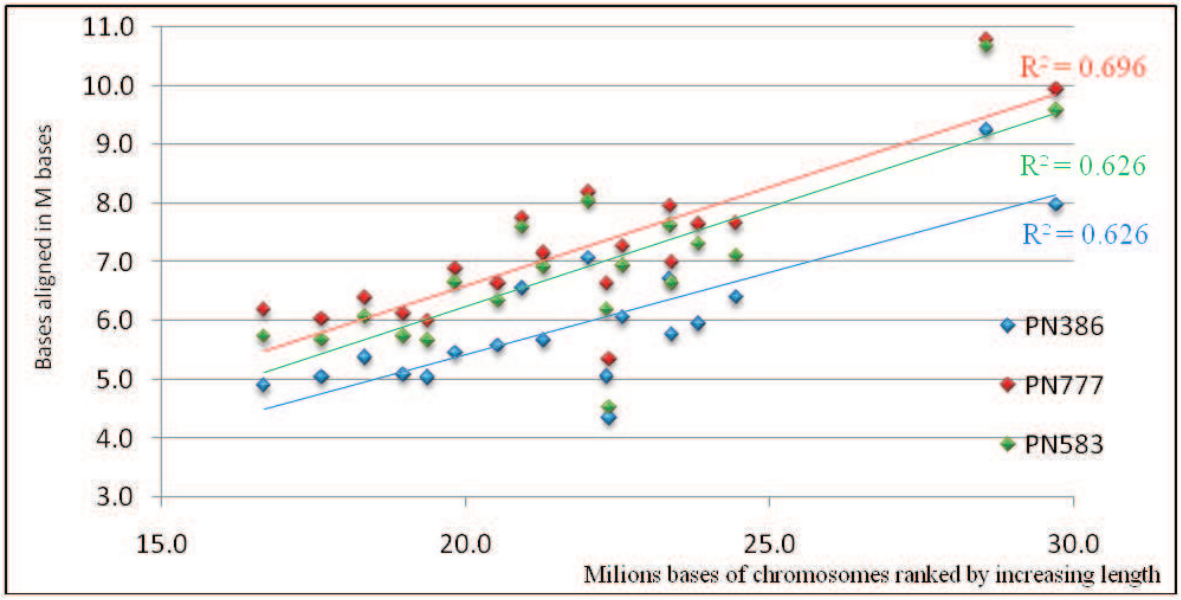
We analyzed sequences of four clones of Pinot noir (PN115, PN386, PN583 and PN777) selected to maximize the phenotypic diversity of this cultivar.

PN115 sequences were downloaded from ncbi database (http://www.ncbi.nlm.nih.gov/) and correspond to published work (Velasco et al., 2007). PN386, PN583, PN777 sequences were obtained by 454 sequencing methodology. These four sets of sequences were aligned on reference sequence PN40024 (Jaillon et al., 2007). For PN115 a total of 67% of the sequences were aligned with the 3 steps procedure (Table 1). They correspond to single locus regions. Since sequences matching more than one locus were discarded. For the other clones an average of 62% of reads was aligned on the PN40024 sequence (Table 1). This represent a mean coverage of 32% of PN40024 sequence at 1.00 fold genome coverage (base count) but only 0.3% at 6.00 fold genome coverage (Table 2).

Among unaligned sequences, only 8% of the reads did not match any known reference sequences of PN40024 (Table 1). These sequences may be either unknown repeated elements, unassembled regions of PN40024 or due to a contamination not reported in any database. The remaining unaligned reads which corresponding to paralog (12%) and repeat sequences (13%), were not retained due to multiple possible localizations on the reference sequence. Reads alignment quality was estimated using an alignment quality score (ranging from 0 to 90) (Ewing and Green, 1998), 90% of the aligned sequences have a quality score higher than 60 (see Supplementary Figure 1).

Comparison with the reference genome

We compared several criteria (percentage of exons, GC, CpG and CnG among the aligned sequences) between clones and PN40024 and no difference were observed (Table 3). The number of aligned bases on each chromosome was proportional to their length (R^2 >0.62, see Supplementary Figure 2). However, our results indicate that read distribution along the



Supplementary Figure 2: Validation of random distribution of aligned reads. The estimated random correlation between the number of aligned reads and the length of the chromosome was tested using Pearson's correlation (P-value < 0.05).

chromosomes was non random and some regions were consistently excluded from alignment (see Figure 1 for an example on chromosome 1). Low-alignment regions showed over-representation of repeat elements in some areas, particularly at the centromere assumed location. There is a significant negative correlation between the number of aligned sequences and the number of repeat elements annotated in the reference sequences (correlation coefficient <-0.25 and p-value <0.01).

Polymorphism calling

In order to eliminate any risk of false positive polymorphism detection from clones sequenced by 454 methodology, we choose to analyze and call polymorphisms only from sequenced regions at 6.00 fold genome coverage (least 6 independent reads should be aligned at each base pair). Moreover, for the polymorphic positions the minor allele should be present in at least 30% of the independent sequences. Because of the absence of common regions between the 3 sequenced clones at 6.00 fold genome coverage we compared each sequenced clone with only PN115 which is a true clone of Pinot noir contrary to PN40024. In total, the sum of the sequences shared by one of the 3 clones and PN115 represents 4.5 Mb (around 1% of grape genome) at 6 fold genome coverage (Table 2). We detected no SSR, but 19 SNPs, 6 indels and 147 sites with a polymorphic insertion of mobile elements (Figure 2A and Supplementary Table 1) representing a mean of 1.6 (+/- 1.0) SNPs, 5.1 (+/- 2.7) indels and 35.2 (+/- 7.2) mobile elements per Mb (Figure 3). Among these putative polymorphisms, 1 indel, 3 SNPs and 19 sites of mobile elements insertion per Mb were localized in genes (predicted from the reference genome -19 March 2010 version- ; Supplementary Table 2). Polymorphisms were well distributed throughout the genome （Figure 2B）.

Distribution and dynamics of mobile elements

Identification of mobile elements

In the partial sequences of clones PN386, PN583 and PN777, we searched for the different mobile elements known in grape (Jurka et al., 2000; Jurka et al., 2005). Among the 107 known mobile elements in grape, 62 have generated at least one insertion polymorphism (see Supplementary Table 3). Polymorphic elements belong to either class I (72%) or class II (23%) mobile elements. The most abundant ones in sequenced clones were LINES retrotransposons, followed by Gypsy and Copia-like elements. However, Gypsy famiy was

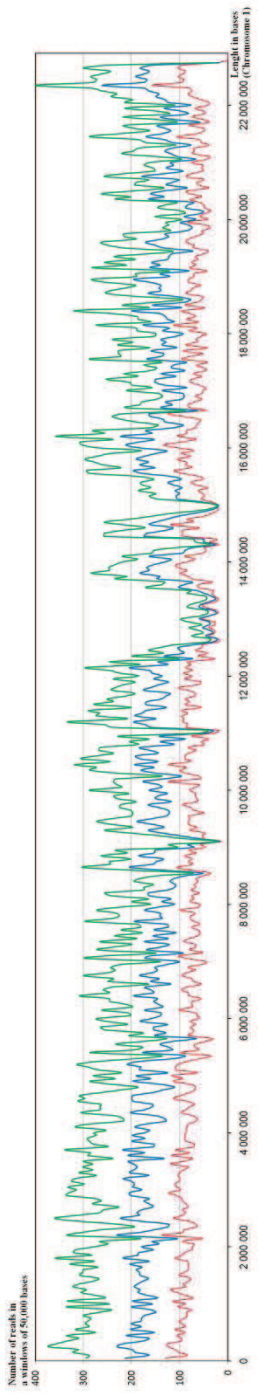


Figure 1: To test the random distribution of reads, three runs were sequentially aligned. The first 454 run was aligned (red line) on chromosome 1. Then both first and second runs were aligned together (blue line), and finally all three runs (green line) were aligned on the chromosome. The insufficiently covered region around 13 Mb in chromosome 1 corresponds to the centromere.

the most elements which generate insertion polymorphisms between clones studies (Supplementary Table 3).

Selection of mobile elements and confirmation of their insertion polymorphism

We selected for detailed analyses four representative mobile elements among class I LTR transposable elements: *Gret-1*, *Copia-10*, *Gypsy-19* and *Cauliv-1*. These four elements have very different copy numbers and polymorphic sites in the partial sequenced of the clones: *Gret-1* displayed 64 copies with 5 polymorphic sites; *Copia-10*, 1273 copies with 4 polymorphic sites; *Gypsy-19*, 564 copies with 3 polymorphic sites and *Cauliv-1* 1065 copies with 2 polymorphic sites (Supplementary Table 3).

To confirm polymorphism due to these mobile elements we performed a S-SAP (Waugh et al., 1997) analysis based on their specific sequences on the 60 Pinot clones registered in France including PN115, PN386, PN583, PN777. We found a total of 134 polymorphic bands (37% of total scored band) among all clones and each clone displayed a specific pattern for these four elements as illustrated in the phenetic tree based on Nei and Li distance matrix (Nei and Li, 1979) from presence/absence of the bands (Figure 4). For the four clones studied in detail (PN115, PN386, PN583, PN777), we found on average 45 polymorphic bands between any 2 clones (see Supplementary Table 4).

Dynamic of mobile elements

LTR distribution and diversity were analyzed in detail for the four mobile elements selected (*Gret-1, Copia-10, Gypsy-19* and *Cauliv-1*). First, within the entire 454 data set, we identified the major forms of consensus LTR and estimated the representation of each of their major forms in the genome (Table 5). Major forms represented by at least 10 locus with 90% identity. Four LTR consensus were identified for *Gret-1* and *Copia-10*, representing 51% and 36% of total LTRs, whereas only one consensus was identified for *Gypsy-19* and *Cauliv-1*, representing less than 10% of the total number of LTRs (Table 4). Minor LTR forms, too divergent to allow building of LTR consensus sequences, represented respectively 93%, 90%, 64% and 49% of identified LTR in *Cauliv-1*, *Gypsy-19*, *Copia-10* and *Gret-1* (Table 5).

Then we built trees based on sequence homology using the conserved region detected in the LTR sequences of these four elements. This conserved region contains the integrase sequence in the 3' LTR (Wicker et al., 2007) (Figure 5). Results for *Gret-1* showed a typical

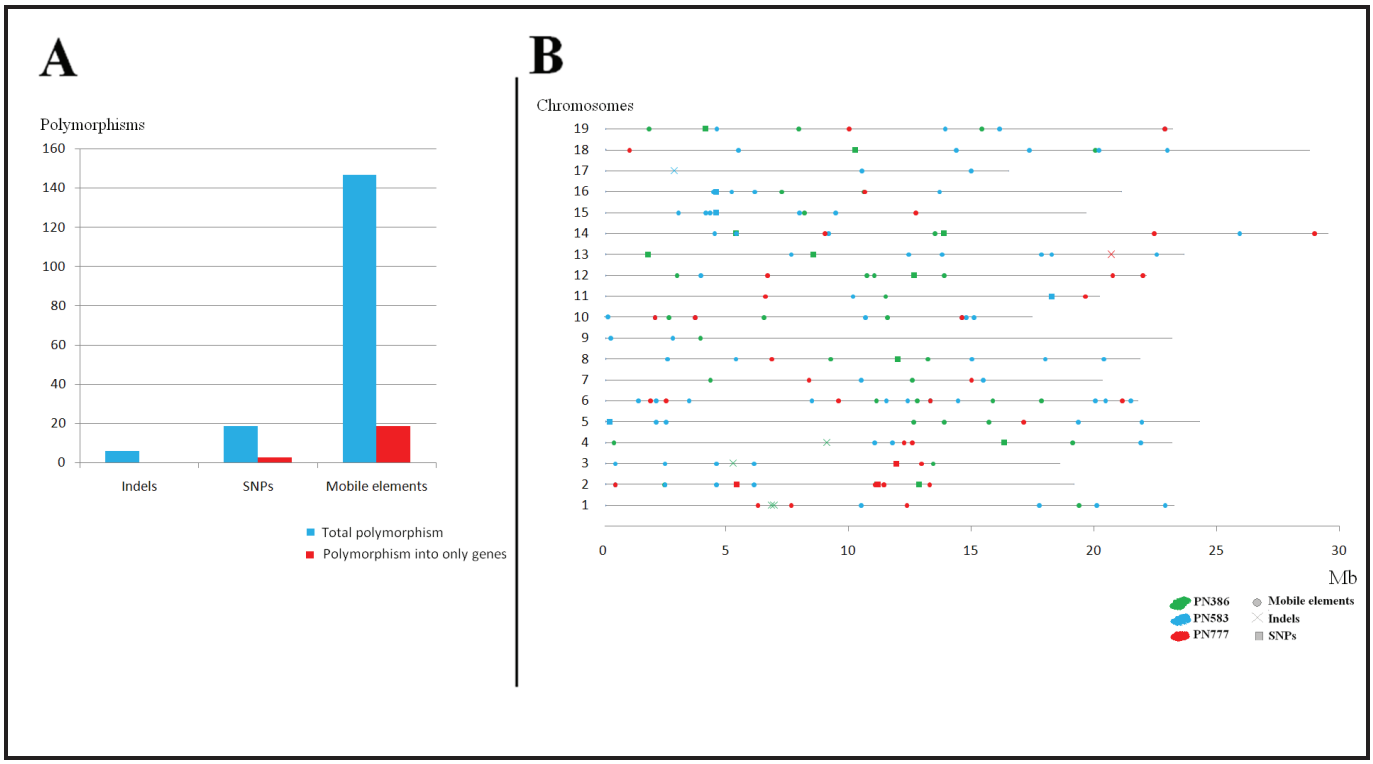


Figure 2: Results from polymorphism call. A) Number of polymorphisms detected in all clones. Numbers of SSR, SNP, indel and mobile element polymorphisms between each pair of clones in regions of 6X coverage only. B) Map of polymorphism in genes between clone PN115 and clones PN386, PN583, PN777. All types of polymorphisms (SNPs, indels, mobile elements) detected between PN115 and partially 454-sequenced (6X) clones (green, blue, red for PN386, PN583, PN777 respectively). SNPs, indels and mobile elements are represented by crosses, squares, and diamonds respectively.

pattern of recent activity with several copies of very homologous sequences. No such patterns were obtained for *Copia-10*, *Gypsy-19* and *Cauliv-1* (Supplementary Figures 4, 5 and 6).

Discussion

The present work represents the first genome-wide analysis of polymorphism among grape clones without *a priori* in an attempt to identify all the molecular polymorphisms involved in somatic mutations. Four Pinot noir clones (PN115, PN386, PN583 and PN777) were selected for their distinct phenotypic characteristics (for example yield or sugar content (Boursiquot et al., 2007)). The clonal selection was performed making prospection in old vineyards, clone PN115, PN386, PN583 and PN777 were selected in different fields in Bourgogne (France) in 1971, 1975, 1978 and 1981 respectively. At this time wood was collected from one particular plant in the field. For each clone history of this plant or of the vineyard was by consequence not available and it is impossible to date the time of divergence between clones. Interestingly, although we have revealed SNPs and indels in this study, the most important mutational events in the context of vegetative propagation were however the insertion polymorphisms generated by mobile elements. Progress in sequencing methods allowed to access to a part of the genome at a total cost and in a time span that were unachievable just a few years ago (Nordborg and Weigel, 2008).

Partial sequencing of Pinot clone genomes

We chose to work on Pinot, one of the most diverse cultivars in term of morphology. An average of 62% of the reads obtained by 454 methodology was aligned at a single locus on the reference sequence and 25% of the reads were not consider because they matched at more than one locus. Our results are similar to those obtained in *Vitis* by Myles *et al.* (Myles et al., 2010). The grape genome is an ancient hexaploid genome (Jaillon et al., 2007) and has many paralogous regions that complicate mapping, particularly for short reads. This is an another reason why we preferred the 454 methodology to any other.

Clone sequenced by 454 methodology (PN386, PN583, PN777) were compared with the PN115 sequence produced by Velasco *et al* (Velasco et al., 2007) which corresponds to assembly with a mean at 6.4 fold genome coverage. In order to perform this comparison, we have aligned all sequences on the reference sequence (PN40024).

SNPs	PN115
PN777	4
PN583	4
PN386	11
Total	19

Mobile el.	PN115
PN777	34
PN583	75
PN386	35
Total	147

IN/DEL	PN115
PN777	4
PN583	1
PN386	1
Total	6

Supplementary Table 1: Detail of polymorphisms detected among clones (SNPs, Indels and Mobile elements) with a depth greater than 6X and a base alignment quality score of more than 60.

S-SAP	PN777	PN583	PN386	PN115
PN777		40	47	41
PN583			58	35
PN386				51
PN15				

Supplementary Table 2: Results of S-SAP for 4 mobile elements analyzed in detail (*Caul-1*, *Gret-1*, *Copia 10*, *Gypsy 19*). Number of polymorphism bands detected between 2 clones generate by 4 mobile elements analyzed.

The random distribution of reads obtained with the 454 method enabled access to a representative part of the grape genome. All chromosomes were covered proportionally to their length, and percentages of GC, CpG and CnG and exon composition were similar between 454 sequences and the reference genome (Jaillon et al., 2007). Major parts of the chromosome regions were easily sequenced and aligned. Only regions containing many repeat elements such as centromere, telomere, and satellite regions were difficult to analyze using this re-sequencing protocol.

Identification of dynamic events involved in somatic genome evolution

We searched for molecular polymorphism among grape clones in order to identify the most significant and dynamic elements involved in vegetative (or somatic) evolution. To limit false positives, only bases sequenced at least six times (corresponding to mean coverage depth of the PN115 sequences (Velasco et al., 2007)) and with alignment quality scores higher than 60 were considered, conditions that have already been used in similar studies (Atanur et al., 2010; Gore et al., 2009; Sabot et al., 2011). Regions shared by PN115 and at least one of the other sequenced clones at 6.00 fold genome coverage represented a total size of 4.5 Mb (approx. 1% of the genome).

Until now, previous studies of clonal diversity, mainly focused on SSRs and AFLP markers, enabled only limited identification of clones (Hocquigny et al., 2004; Imazio et al., 2002; Moncada et al., 2006; Pelsy et al., 2010). Although they present a quite low mutation rate, both SNPs and indel have been identified in our studies and are therefore potential markers to study clonal diversity. The related polymorphism rate is however quite low, since we found 1.6 SNPs and 5.1 indels per Mb, while polymorphism between cultivars can be as high as 20 000 SNP per Mb (Le Cunff et al., 2008). Although they are less abundant than mobile elements, SNPs are known to generate polymorphism when they are located in genes. As an example, one SNP modification in the *VvGAI-1* gene of a Pinot meunier clone resulted in a dwarf phenotype (Thornsberry et al., 2001). In the present study, one SNP between PN777 and PN115, is located in one exon and generates a non-synonymous mutation (Supplementary Table 4). This candidate gene could be associated with phenotypic differences and, considering the low cost of the analysis, one can suggest that clone and/or somatic mutant sequencing might be an interesting way to identify candidate genes linked to grape polymorphism.

Position		Genes	
Mobiles elements			Mobiles elements
chr1	22892753	GSVIVT01001135001	Gypsy22
chr1	7660970	GSVIVT01013782001	Copia1
chr10	2687291	GSVIVT01021220001	Copia1
chr10	15104708	GSVIVT01026261001	VHARB
chr12	6704871	GSVIVT01030556001	Copia23
chr12	21987082	GSVIVT01023147001	VLINE3
chr13	7667083	GSVIVT01034686001	Gypsy12
chr19	1861065	GSVIVT01014241001	Gypsy9
chr2	499949	GSVIVT01019417001	Harbinger_1
chr2	6150229	GSVIVT01013259001	Copia10
chr5	15713361	GSVIVT01020995001	VHARB4
chr5	2568893	GSVIVT01017678001	Gypsy17
chr5	21944397	GSVIVT01010735001	Gypsy22
chr6	13331536	GSVIVT01037457001	Copia3
chr6	14461205	GSVIVT01037393001	Gypsy12
chr6	1440555	GSVIVT01025364001	VLINE1
chr8	13235871	GSVIVT01025657001	Copia22
chr8	2619320	GSVIVT01029978001	Gypsy9
chr8	20392045	GSVIVT01033475001	Gypsy22
SNPs			Located
chr5	267707	GSVIVT01024255001	intron
chr12	12667503	GSVIVT01011544001	intron
chr13	8573986	GSVIVT01034732001	intron
chr18	10288579	GSVIVT01023526001	exon - non synonimous - putative peptidase
chr19	4168276	GSVIVT01014467001	intron
Indels			
chr1	6979210	GSVIVT01000570001	

Supplementary Table 4: Polymorphisms located in genes.

The major cause of somatic polymorphisms were insertion polymorphisms caused by mobile elements since 147 events were observed (35.2 per Mb). Such great extents of mobile elements polymorphism strongly suggest somaclonal transcriptional activation. Mobile elements are known to generate a substantial number of mutations that can impact gene expression and genome size, while sequence duplications can also be responsible for new gene functions (Feschotte et al., 2002; Kazazian, 2004; Wicker et al., 2007). In grape, variation of grape berry color for example was due to the insertion of the *Gret-1* element into the *VvMybA-1* promoter (Kobayashi et al., 2005). In our study, 19 out of 147 events involving mobile elements are found in genes. These specific elements could be used in the future with S-SAP or other protocols to study clonal diversity.

This level of polymorphism generated by mobile elements is high. Validations on other samples are presently in progress on genome wide analysis of clonal variation. It will allow comparisons with diversity at cultivar level as wel. Since no other work has been reported comparison is impossible. Nevertheless, S-SAP analysis using 4 elements (*Gret-1, Copia-10, Gypsy-19, Cauliv-1*) also revealed high insertion polymorphisms generated by mobile elements: 30% of total bands were polymorphic between clones. Moisy *et al*., (Moisy et al., 2008) studying distribution of mobile elements in 7 cultivars using S-SAP observed that 80% of the bands were polymorphic between cultivars showing high polymorphism between cultivars.

Dynamics of mobile elements linked to vegetative multiplication

For all partially analyzed genomes, we determined the number of copies of each mobile element (Supplementary Table 3). The LINES retrotransposon family was the most widely represented (5 LINES among the 6 most abundant elements) followed by Gypsy and Copia-like elements. The same result was obtained in the reference genome, with 75% of repeat elements corresponding to LINES members (Jaillon et al., 2007). Activity of Gypsy family elements is known to generate high polymorphism in plants (Feschotte et al., 2002) and indeed, although they were less numerous than LINES elements, Gypsy elements showed higher polymorphism than LINEs.

We analyzed LTR distribution and diversity in detail for the four mobile elements (*Gret-1, Copia-10, Gypsy-19* and *Cauliv-1*) and identified for each element several consensus LTR which could be correlated to mobile elements activity. In fact, the more frequent

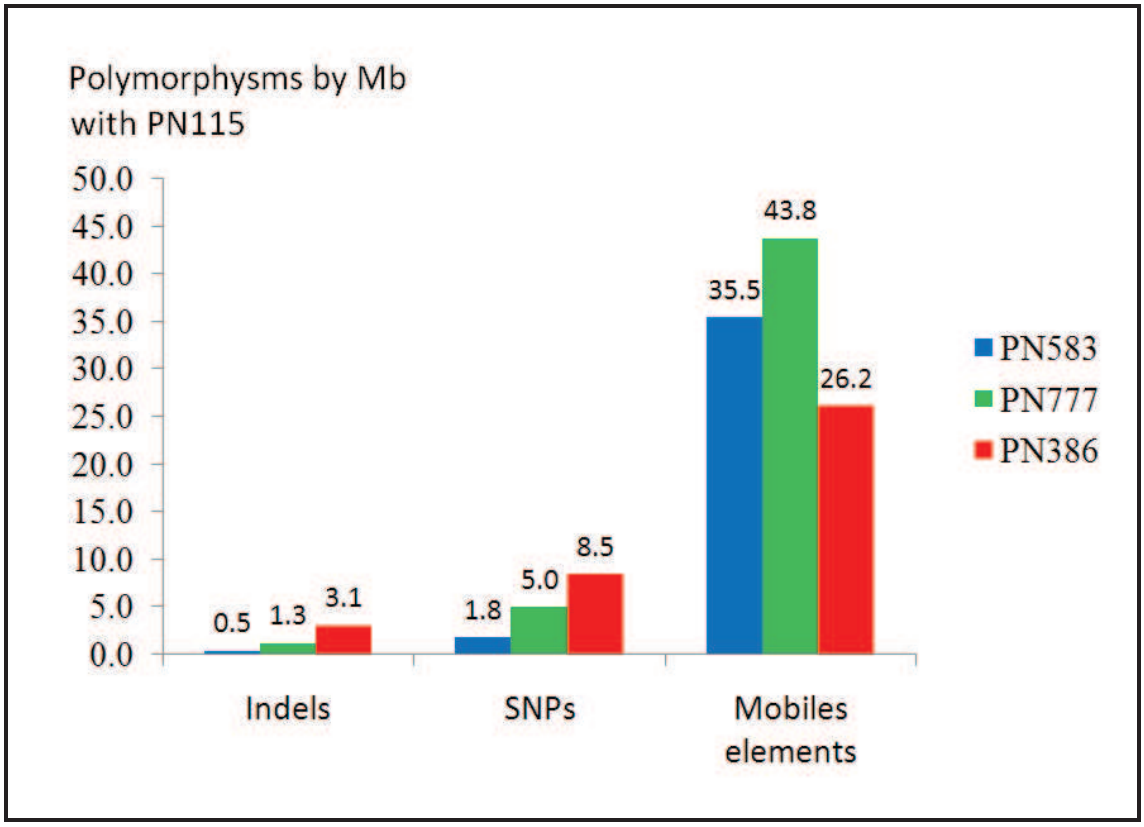


Figure 3: Number of SSR, SNP, indel and mobile element polymorphisms with PN115 per 1 Mb of genome sequence for each clone.

representation of major forms over minor forms for one element suggests a high level of recent activity. Interestingly, in our study, mobile elements ranked in the same order when classified by their percentage of major forms or by their number of polymorphism insertions, conforting analysis accuracy (Supplementary Table 3 and Table 4). *Gret-1* had the lowest proportion of minority forms and generated most of the insertion polymorphism in all partially analyzed genomes. In contrast, *Cauliv-1* had the highest proportion of minority forms and generated the lowest level of insertion polymorphism among the 4 studied elements.

Figure 5 shows the pattern displayed by *Gret-1* with similar LTR sequences that had no time to diverge. In the last years, studies have shown that *Gret-1* is a "recent" mobile element (Llorens et al., 2009; Moisy et al., 2008) with reportedly recent activity since *Gret-1* insertion into the *VVMybA1* color regulating gene is believed to have occurred after grape domestication some 7000 years ago (Fournier-Level et al., 2010).

A list of potential markers

The S-SAP approach has been used to analyse clonal diversity but with very contrasting results according to the mobile elements tested. Wegscheider *et al*. (Wegscheider et al., 2009) used universal retrotransposon-based primers and revealed polymorphism among five Pinot clones. But Verriès *et al*., (Verries et al., 2000) using *Vine-1* based primers, failed to reveal any variation among six Pinot clones. A wider choice of mobile elements which can be used as markers in clone diversity studies could therefore be very appropriate and the list of mobile elements presented in this paper may thus help the grapevine genetics community in the selection of efficient markers. We tested four of these elements with a high level of insertion polymorphism (*Gret-1, Copia-10, Gypsy-19* and *Cauliv-1*) in Pinot clones registered in France. Each clone displayed a specific pattern for these elements (Figure 4), thus confirming the high level of insertion polymorphism they could have generated by transposition activity. Although this was not the aim of our study, these elements might be used to study diversity in Pinot and other grape cultivars as all four Cabernet Sauvignon clones studied here *(CS15, CS191, CS216, CS416)* also displayed a specific pattern for these mobile elements (Figure 4). Caution should however be exercised in the use of S-SAP as this method might be hindered

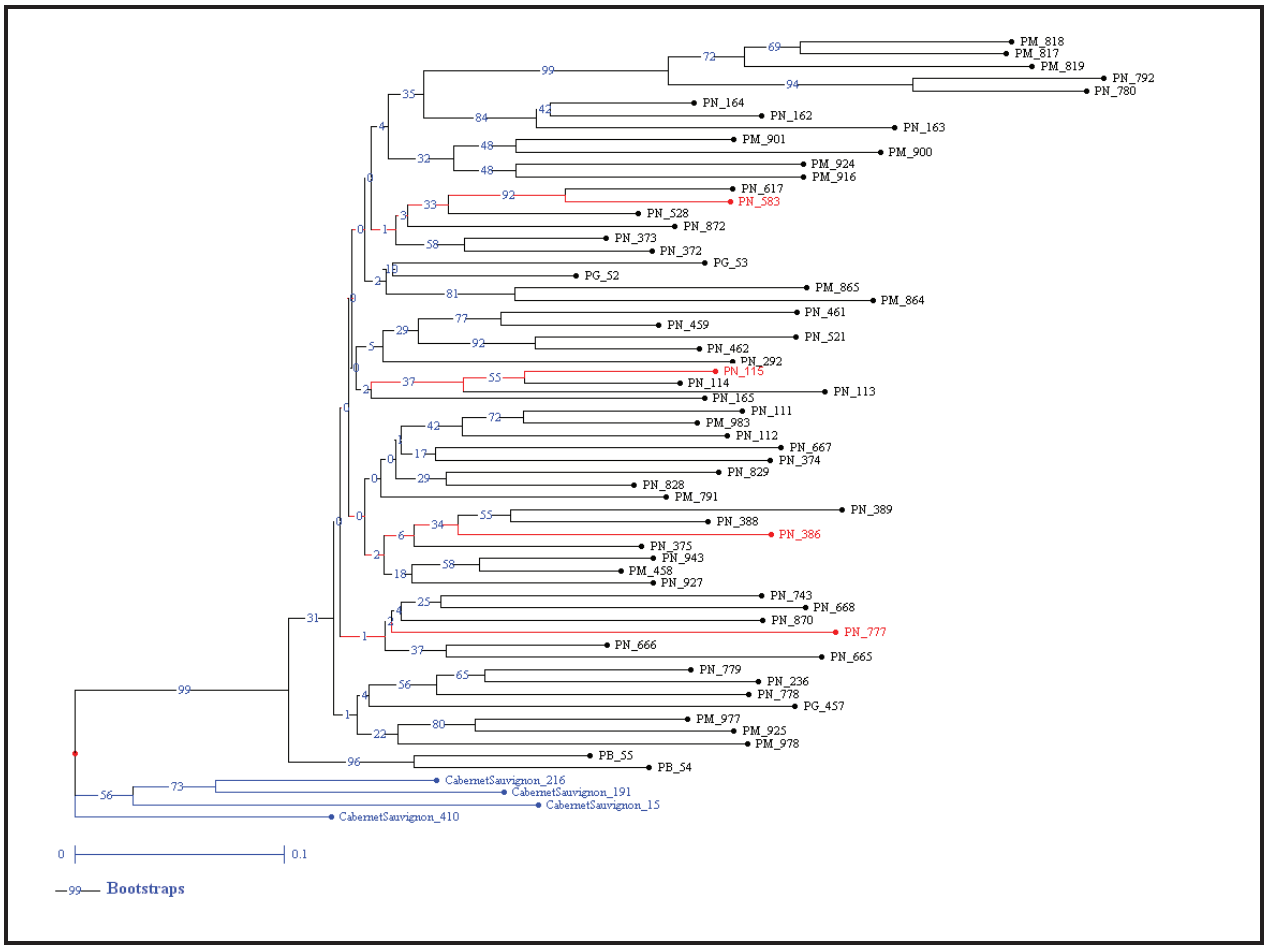


Figure 4: Phylogenetic tree of all registered Pinot clones based on S-SAP with 4 mobile elements. S-SAP performed with 4 selected mobile elements (Gret-1, Copia-10, Gypsy-19 and Cauliv-1). All analyzed clones have a specific pattern for these elements. 60 Pinot clones (PN= Pinot noir (40) ; PM=Pinot meunier (15) ; PG= Pinot gris (3) ; PB= Pinot blanc (2)) and 4 Cabernet-Sauvignon clones were analyzed.

due to high mobile element activity. Markers base specific locus should therefore be preferred.

Conclusion

Genome-wide comparison of spontaneous grape clones enabled the first study of the molecular polymorphisms generated along vegetative propagation at whole genome scale. Although a small number of SNP and indel events were also observed, mobile elements were involved in most polymorphisms. Gypsy-like elements being were the most polymorphic ones. This study identified 172 polymorphic sites in a cumulative analysis of 4.5 Mb of the grape genome, which represent a higher polymorphism level than initially expected for vegetative propagation material. Additional analyses are now underway in order to analyze a larger part of the genome of the clones already studied as well as new clones and work clones of other cultivars to confirm our results.

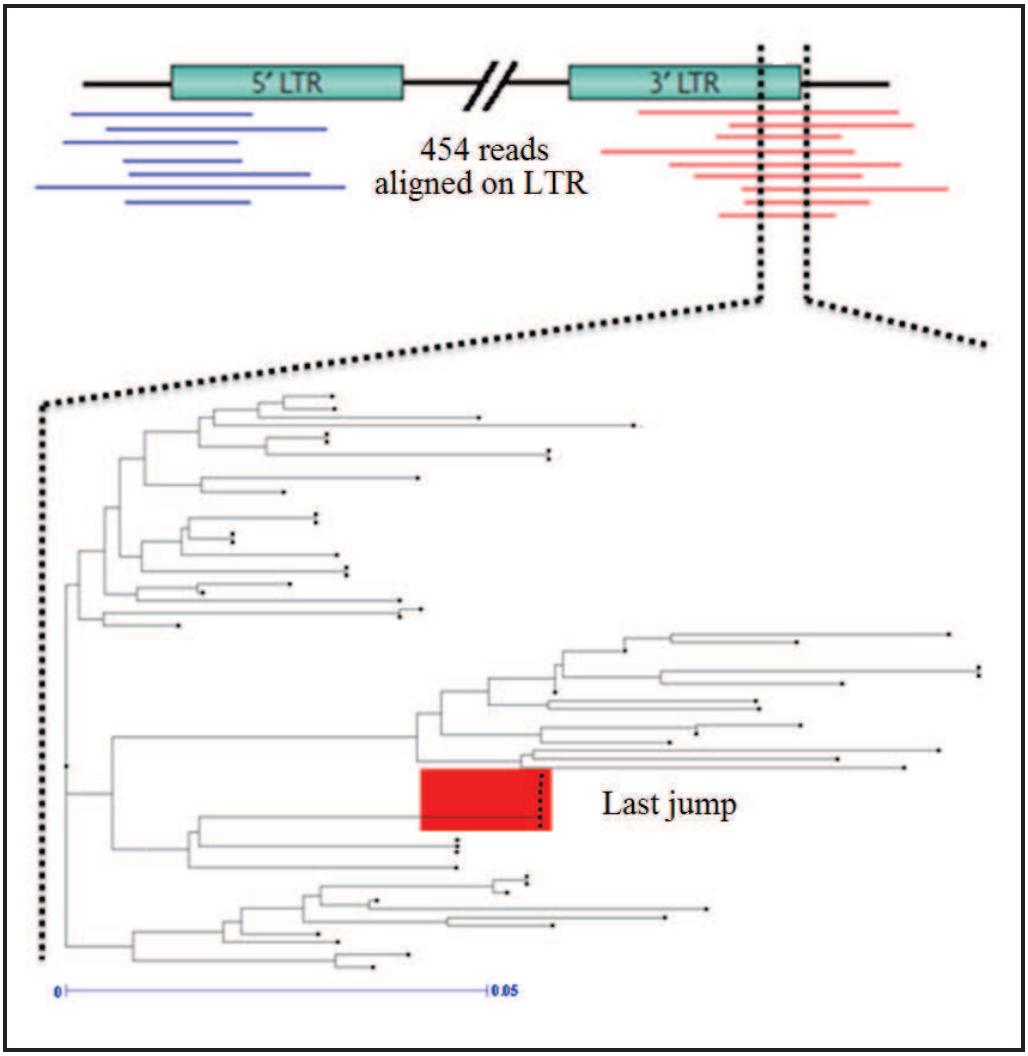


Figure 5: Dynamics of the four mobile elements.

The consensus region used to build the tree for *Gret-1* mobile element is indicated by the dashed line. On the tree, the group of similar sequences (circled in red) suggests recent activity of *Gret-1*.

Material and methods

Plant material and DNA extraction

Three clones of *Vitis vinifera L.* cultivar Pinot noir n° ENTAV-INRA® 386 (PN386), 583 (PN583) and 777 (PN777), grown at the Espiguette repository, were selected for maximum phenotypic diversity. These Pinot clones were selected by ENTAV-INRA® in Bourgogne (France) in 1975, 1978 and 1981 for PN386, PN583 and PN777 respectively. PN777 is the clone producing the highest quality wine than PN583 and PN386 (Boursiquot et al., 2007). We harvested 5 g of young leaves for nuclear DNA extraction using the NGS method previously described (Carrier et al., 2011). S-SAP studies were performed on the registered Pinot clones (2 Pinot blanc, 3 Pinot gris, 15 Pinot meunier and 40 Pinot noir) grown in the Espiguette collection. DNA extraction was performed with Qiagen MaxiQKit® according to the manufactory instructions.

Sequencing samples of PN386, PN583 and PN777 genomes

Approximately 5 µg of nuclear DNA were used for 454 GS-FLX sequencing as previously described (Margulies et al., 2005) at the Genotoul platform (INRA Toulouse Midi-Pyrénées). The data is available from NCBI (FastQ files: SRX098092 for PN386; SRX098091 for PN583 and SRX098090 for PN777). Reads produced using 454 methodology were analyzed with FastQC software (v0.6) developed by Simon Andrews at the Babraham Institute (www.bioinformatics.bbsrc.ac.uk) to validate run quality (sequence number, mean sequence length etc.). We obtained approx. 350 Mb (330 – 378 Mb) per run, corresponding to approx. one million reads with an average length of 355 bases (Table 5). In terms of base quantity, PN583 was the best run, while both PN777 and PN386 were slightly better in terms of quality (quality score on Phred Sanger graduation (Ledergerber and Dessimoz, 2010)). Quality decreased proportionally with read length (Supplementary Figure 3). Duplicated sequences generated by EmPCR bias represented an average of 4974 reads per run (0.58%). There were no overrepresented sequences per run and a very low percentage of contamination by other organisms (132 reads per run on average).

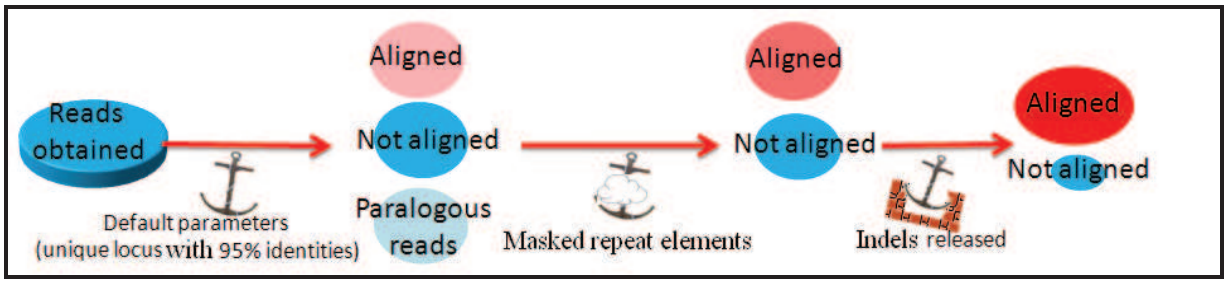


Figure 6: Summary of the alignment method used in the present study. Alignment was accomplished in three successive steps:

i) The first alignment used Mosaik with default parameters for 454 GS-FLX Titanium: 95% alignment homology in the sequences;

ii) Reads not aligned in the first step and that were not paralogs were then filtered with RepeatMasker software. Reads with less than 90% homology with repeat elements were aligned by Mosaik with default parameters

iii) For reads not aligned in the second step, a third alignment was performed using a gap parameter fixed at a minimum (0.1 gap open and extensive penalties).

Aligning PN115, PN386, PN583 and PN777 with the reference genome (PN40024-12X)

We used the Hash-based alignment methods incorporated in the MosaikAssembler tool v1.0 (Wan-Ping Lee and Michael Strömberg, available at bioinformatics.bc.edu/marthlab/). The data set was composed of reads obtained by 454 methodology and PN115 sequences downloaded from NCBI, (Project ID: 18357, www.ncbi.nlm.nih.gov/) (Velasco et al., 2007). In order to avoid a bias of sequence alignment between the clones studied, the contigs and scaffolds from the PN115 sequences were sheared *in silico* to be considered as data from 454 sequences (size 1000 bases), assuming each nucleotide with optimal quality score.

Sequences of each sample were aligned on the reference genome sequence (PN40024, 12X version (12-Feb 2010)) in three steps: i) alignment of single reads that shared 95% homology with PN40024, ii) unmatched reads were masked for repeat elements and aligned if at least 150 bases were not masked, iii) for the remaining sequences, relaxed stringency was applied with no impact of the gap parameter (Figure 6 and Table 1) (For details on the alignment method, see supplementary data). The origin of non-aligned reads was identified as : i) reads composed of 90% repeat sequences; ii) reads aligned at two loci or more, paralogous reads; iii) reads of cytoplasmic origin (> 90% sequence identity with *Vitis vinifera* chloroplast: NC 007957 or mitochondrion: NC 007762); iv) contamination reads originating from other organisms known to be present in laboratories (> 90% of identity with *Saccharomyces cerevisiae* S288c (Project ID: 128), *Escherichia coli* 536 (Project ID: 16235), and v) too short (100 pb) or low quality (< Q20) (Mosaik filter) reads (Table 5).

Polymorphism calling

For all polymorphism calling, identification was first performed *in silico* and all polymorphic loci were then validated manually using EagleView (Huang and Marth, 2008). This manual validation was essential for the following reasons: i) the 454 method is known to create some false positives, particularly with homopolymer sequences ii) the parameters we used for the third alignment (gap parameter fixed at a minimum) may also have created some false positives.

All polymorphisms between 2 clones were called with Gigabayes (http://bioinformatics.bc.edu/marthlab/) between two clones. To reduce false-positive rate,

Detail of alignment method

We used Hash-based alignment methods available in the MosaikAssembler tool v1.0 (Wan-Ping Lee and Michael Strömberg, available athttp://bioinformatics.bc.edu/marthlab/). The alignment used the reference sequence available at (http://www.genoscope.cns.fr, unmasked version dated 12-Feb-2010). The data set comprised reads obtained in the 454 and PN115 sequences downloaded from NCBI, Project ID: 18357, http://www.ncbi.nlm.nih.gov/). PN115 sequences were considered like the 454 sequence. The download sequence of PN115 provides consensus sequences produced by (Velasco *et al.* 2007). We considered the optimum quality for this nucleotide sequence (40). Sequences of each sample were aligned on the reference genome sequence (PN40024) in three steps (Figure 1). For the first alignment we used the default parameters provided by MosaikAssembler with 454 GS-FLX Titanium (95% homology with reference sequence and only reads at a single aligned locus). Reads aligned on many loci (paralogous reads) were filtered. In the second alignment, reads that were not aligned in the first step were filtered by RepeatMasker software (Chen 2002) using library by default and adding the grape mobile element data bases (http://www.girinst.org/repbase/) (Jurka *et al.* 2000, Jurka *et al.* 2005). Reads that were masked but contained a minimum of 150 not masked bases were aligned by MosaikAssembler with default parameters. The third alignment was performed using a gap parameter fixed at the minimum (0.1 gap open and extensive penalties). We added a tag in order to identify the origin of sample and alignment steps.

Identification of the origin of unmapped reads was accomplished in five classifications: i) reads composed of 90% repeat sequences; ii) reads mapped at two or more loci, paralogous reads; iii) reads of cytoplasmic origin (< 90% of identity with *Vitisvinifera* chloroplast: NC 007957 or *Vitisvinifera* mitochondrion: NC 007762); iv) Contaminated reads from other organisms known to be present in laboratories (< 90% of identity with (Saccharomyces cerevisiae S288c(Project ID: 128), Escherichia coli 536 (Project ID: 16235),and v) bad quality too short reads (100 pb) or bad quality reads (< Q20) (Mosaik filter).

To confirm the correct representation of our aligned data, we compared some criteria (percentage of GC, CnG, CpG and exons) with the reference genome using home perl script and Blast2.0 algorithm (Altschul *et al.* 1990).

we chose to select polymorphism at a given position, only if a 6.00 fold genome coverage or more was obtained for each clone, and if minority alleles displayed a minimum frequency of 0.3 with an alignment quality score higher than 60 (Ewing and Green, 1998). Polymorphic indels were considered only if they were surrounded by a sequence not localized in the read

terminal region and to limit false positives, none of the reads aligned after the third alignment step was used for indel polymorphism detection. A filter was used with RepeatMasker to identify mobile element-linked polymorphisms (Chen, 2002). Reads composed of a minimum of 150 unmasked bases and a minimum of 100 masked bases were aligned and localized in the reference genome. This polymorphism was called with Gigabayes: indels detected on masked reads were considered as mobile element polymorphisms.

S-SAP was used to validate mobile elements polymorphism as in previously published studies (Knox et al., 2009; Labra et al., 2004; Wegscheider et al., 2009) (for details see supplementary data). Primers for retrotransposons were based on sequenced reads containing the LTR region. We chose the most conserved LTR region to design primers in order to amplify the largest transposition loci. A phenetic tree was based on Nei and Li distance matrix (Nei and Li, 1979) from presence/absence data and was built with Darwin software (Perrier and Jacquemond Collet, 2006) with 1000 permutations (Figure 4).

Studies of mobile elements activity in the clones' genome

Four mobile elements were analyzed in detail (*Copia10, Gret-1, Gypsy-19* and *Cauliv-1*). Each insertion generated by these mobile elements was detected and major forms of these element were detected from consensuses form build using AAARF software (DeBarry et al., 2008) with the following parameters: 10 LTR reads min, 90% identity. LTR homology sequence trees were obtained using the ClustalW algorithm (Thompson et al., 1994) with 1000 permutations and the neighbor-joining method (Saitou and Nei, 1987).

Authors' contributions

GC performed the grape sequencing using the 454 method, analyzed the data and drafted the manuscript. AD helped develop the perl script for data analysis. DL and BO helped perform grape sequencing at the GENOTOUL platform. FS participated in drafting the

Acknowledgements

We are grateful Dr. Anne-Francoise Adam Blondon, Dr. Frédérique Pelsy, and Dr. Franc-Christophe Baurens for discussion in this study. We acknowledge Daphne Goodfellow and Dr Philippe Chatelet for improving the English.

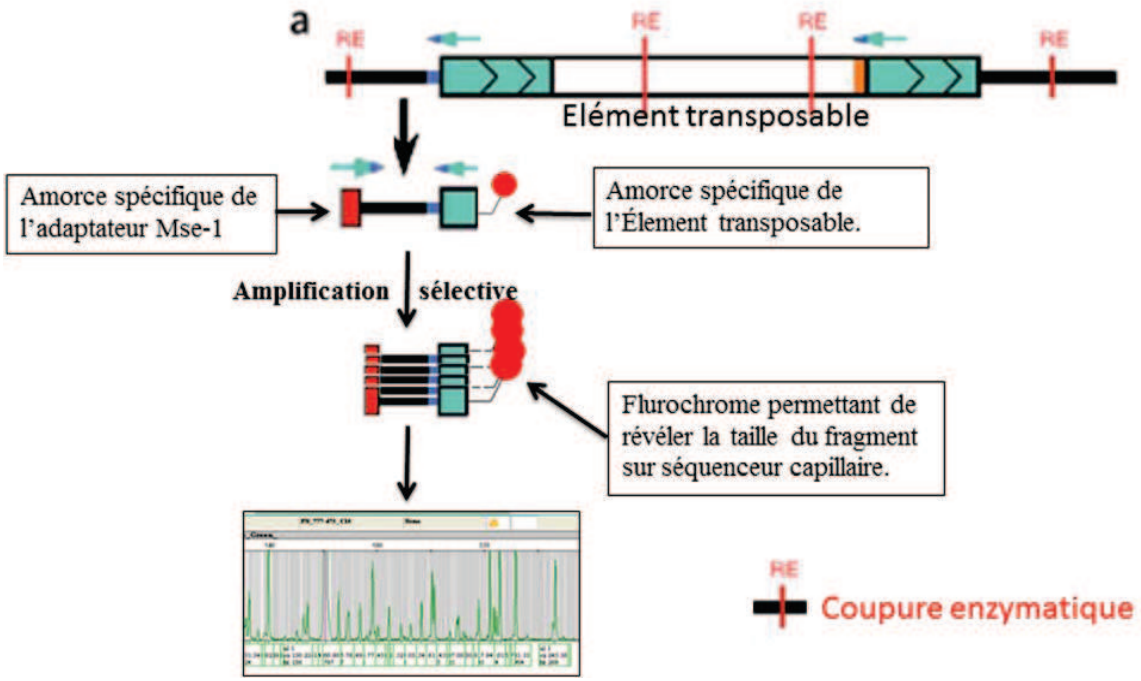
This work was funded by the French Ministry of Research and Higher education and the French Ministry of Food, Agriculture and Fisheries; including a PhD grant from the IFV for GC.

La méthode S-SAP

La S-SAP (Sequence-Specific Amplification Polymorphism) permet d'identifier les multiples insertions d'un élément transposable dans le génome. Elle est basée sur l'amplification de régions spécifiques contenant l'insertion de l'élément transposable pour lequel une amorce a été dessinée.

Dans un premier temps, l'ADN est digéré avec une enzyme de restriction (Mse-1 par exemple). Des adaptateurs spécifiques aux coupures de Mse-1 sont ligués à chaque fragment d'ADN. Une amplification sélective est ensuite réalisée avec une amorce spécifique de l'adaptateur Mse-1 et une amorce spécifique de l'élément transposable étudié (Cf. Figure). Chaque fragment amplifié est caractérisé par sa taille. Pour les éléments présentant un grand nombre de copies dans le génome, des bases dites sélectives sont ajoutées à l'amorce Mse-1 permettant de ne pas saturer le signal d'amplification. Généralement nous ajoutons 4 bases sélectives à l'amorce Mse-1. La taille des fragments amplifiés est ensuite révélée par électrophorèse par exemple à l'aide d'un séquenceur à capillaire.

Pour plus de détails voir la revue de Naeem et al., (2007)



5-Etude de la diversité du Pinot à l'aide de quatre éléments transposables.

Une analyse S-SAP réalisée à partir des quatre éléments transposables étudiés précédemment (*Gret-1, Copia-10, Gypsy-19, Caul-1*; Cf. Chapitre 3, Section 4) a permis d'établir un arbre phylogénétique des 60 clones de Pinot agréés en France et de quatre clones de Cabernet-Sauvignon. Cette analyse a déjà été présentée dans l'article de la section 4. Dans cette partie, nous avons mis en relation la structuration obtenue avec des données phénotypiques et d'origine disponibles. Ces données sont essentiellement issues du catalogue officiel des variétés et clones cultivés en France (Boursiquot *et al.,* 2007). Ces résultats préliminaires, permettront d'établir si la diversité des clones de Pinot agréés est structurée en fonction de leurs caractères ampélographiques majeurs (couleur, villosité) ou de leurs origines géographiques. Dans un second temps, nous avons comparé la diversité des clones de Pinot agréés aux clones de Pinot non agréés maintenus dans les différents conservatoires.

5.1- Structuration de la diversité des clones de Pinot agréés en fonction des variétés

Nous avons étudié un petit nombre de clones de Cabernet-Sauvignon et quatre variétés ou groupes de clones de Pinot: Pinot noir, Pinot blanc, Pinot gris, Meunier. La figure 40 présente la diversité des clones en fonction de leur appartenance à chaque variété. On remarque d'une part la séparation nette en deux branches distinctes des deux cépages Cabernet-Sauvignon et Pinot. Au sein des clones de Pinot, on observe également nettement un regroupement des deux clones de Pinot blanc analysés et leur séparation par rapport aux autres clones. Par contre, il est très difficile de trouver une structuration des clones, respectivement, de Pinot noir, Pinot gris et de Meunier. Ces résultats peuvent être interprétés de deux façons différentes: i) Soit le polymorphisme d'insertion généré par les éléments transposables n'est pas accumulé de façon permanente au cours de la vie du clone et des cycles de multiplication végétative. ii) Soit la généalogie des clones est très complexe. Les clones de Pinot blanc, seraient issus d'un ancêtre commun et donc se regroupent. En revanche, les clones de Meunier et les clones de Pinot gris ne se regroupent pas, les mutations grises et Meunier seraient donc aparus à plusieurs reprises chez des individus différents. Notons que le

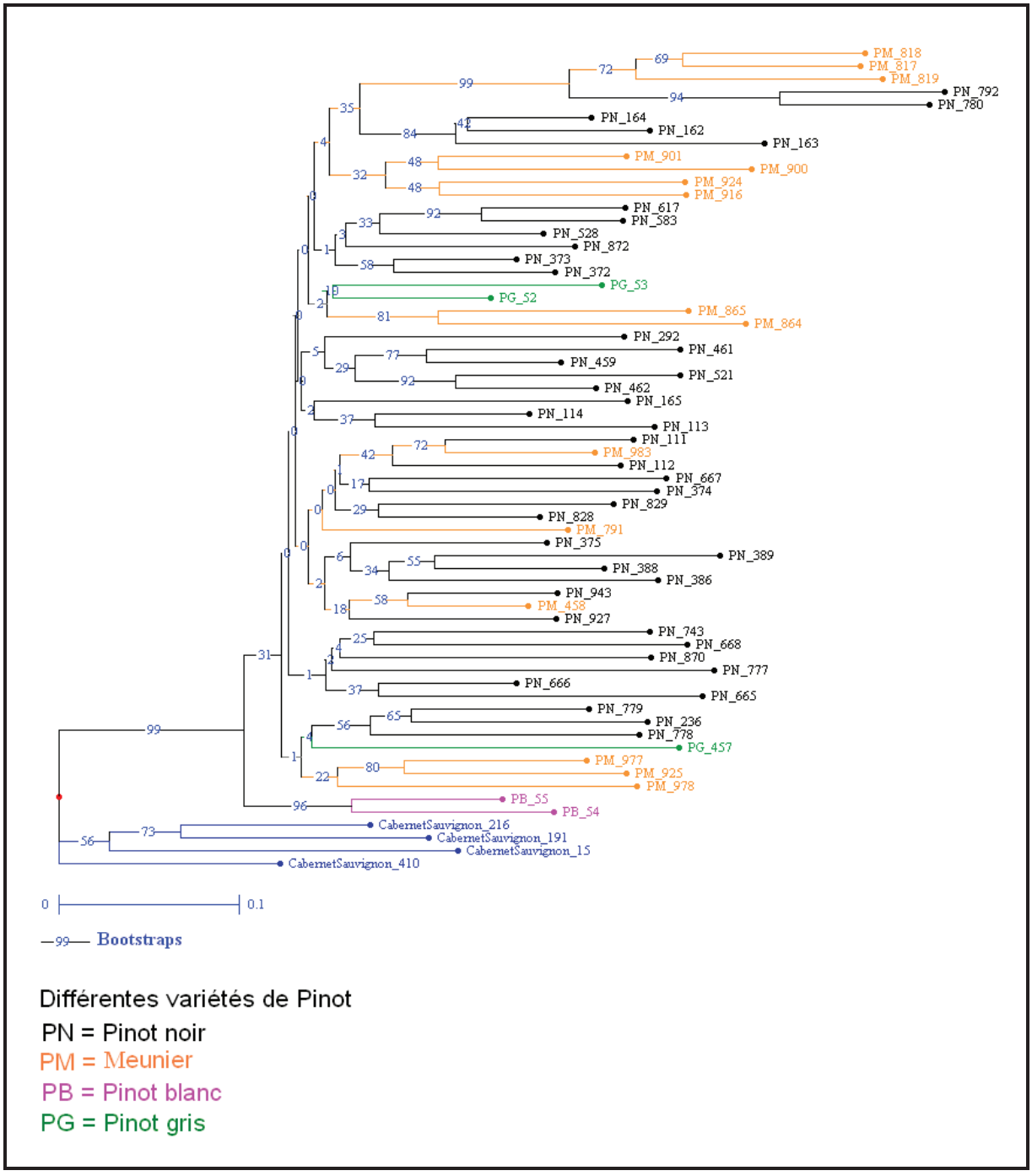


Figure 40 : Arbre phylogénétique obtenu à partir des profils S-SAP de 4 éléments transposables (*Gret-1 ; Copia-10 ; Gypsy-19 ; Cauli-1*) en fonction des caractéristiques phénotypiques spécifiques aux variétés de Pinot.

caractère meunier est connu pour être instable et qu'il n'est pas rare d'assister à des réversions/apparitions de ce caractère (Figure 41). Il est donc cohérent de retrouver des clones de Meunier répartis dans tout l'arbre.

5.2- Structuration de la diversité des clones de Pinot agréés en fonction des variétés et de leur origine géographique

De la même façon, nous avons recherché si l'on pouvait mettre en relation la structuration obtenue et l'origine géographique des clones. On entend par origine géographique la localisation du centre qui a prospecté et mis en place les collections de clones (Boursiquot *et al.,* 2007). La figure 42 présente une distinction des clones de Pinot en fonction de leur origine géographique. En dehors des clones de Pinot blanc issus tous les deux d'Alsace, il n'y a aucune structuration nette des clones en fonction de leur origine géographique. Tout au plus peut-on observer quelques branches de l'arbre préférentiellement composées de clones issus d'un même bassin viticole.

Il ne faut pas oublier que le matériel végétal a beaucoup voyagé en France et il est également possible que les clones présents dans les conservatoires régionaux aient été prospectés en dehors de régions viticoles. Tout comme pour l'analyse précédente, il est également possible que les marqueurs S-SAP ne soient pas suffisamment stables pour effectuer une telle analyse.

5.3-Comparaison de la diversité des clones de Pinot agréés et des clones de Pinot présents dans les conservatoires

Les conservatoires de clones de Pinot non agréés comptent environ 300 individus prospectés et sélectionnés pour leur potentiel agronomique. Les conservatoires sont la source des futurs clones agréés, ces individus ayant déjà subi une première étape de sélection. Nous avons comparé la diversité génétique entres les clones de Pinot agréés et une partie de ceux contenus dans les conservatoires. Une approche S-SAP à partir de l'élément transposable *Gret-1* a été réalisée chez les 60 clones agréés et les 254 clones contenus dans les

Figure 41 : Réversion sectorielle de la mutation meunier (forte villosité) (Photographie J.M Boursiquot).

conservatoires. Nous avons sélectionné *Gret-1* parce que c'est un des éléments transposables les plus récent (Cf Chapitre 3, Section 5 et Moisy *et al*., 2008) donc probablement avec une forte activité de transposition.

Les clones analysés se regroupent sur cinq branches majeures de l'arbre (Figure 43). Les clones agréés se retrouvent quant à eux essentiellement sur deux des branches où ils se regroupent avec très peu de clones non agrées présents dans les conservatoires. La diversité présente aujourd'hui dans les conservatoires est donc plus importante et relativement indépendante de celle observée chez les clones agréés.

6-Synthèse sur le polymorphisme moléculaire chez le Pinot

Le polymorphisme clonal chez le Pinot noir a été identifié et quantifié en comparant une partie du génome de différents clones. Pour la première fois, des mutations ponctuelles (SNPs 5,1 par Mb et indels 1,6 par Mb) ont été détectées entre les clones. Ces mutations sont cependant minoritaires par rapport au polymorphisme d'insertion généré par les éléments transposables (35,1 par Mb). L'activité des éléments transposables est aussi la source majoritaire du polymorphisme génétique entre les clones.

6.1-Deux dynamiques mutationnelles à l'origine de la variation clonale

Une dynamique différente entre les mutations ponctuelles et les éléments transposables a pu être observée. L'apparition des mutations de types SNPs est relativement faible et constante par unité de temps (estimée à 10^{-7} par mitose (Cloutier *et al.*, 2003)) alors que l'activité des éléments transposables est très fluctuante et dépend de l'état de stress de la plante (Grandbastien, 1998). Nous n'avons aucune référence sur l'origine de l'apparition des SNPs et indels alors que Moisy (2008) a pu mettre en relation les blessures causées par la taille de la vigne et l'activité de transposition de certains éléments transposables. Les conditions culturales, particulièrement stressantes pour la vigne favorisent l'activité des éléments transposables et participent à un accroissement de la diversité clonale.

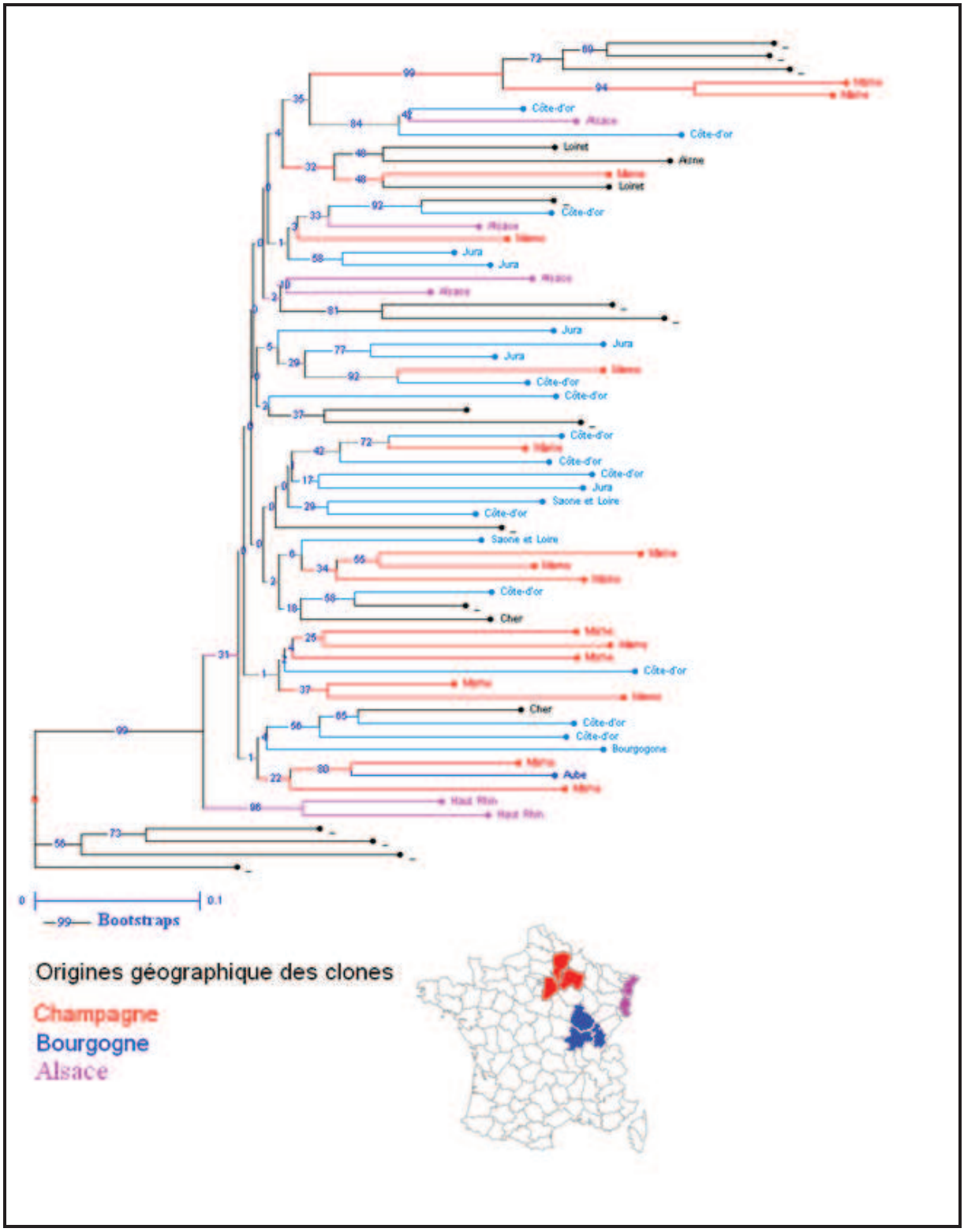


Figure 42 : Arbre phylogénétique obtenu à partir des profils S-SAP de 4 éléments transposables (*Gret-1 ; Copia-10 ; Gypsy-19 ; Cauli-1*) en fonction de l'origine géographique des clones de Pinot.

6.2-Polymorphisme d'insertion généré par les éléments transposables connus chez la vigne

Une liste des éléments transposables connus chez la vigne entrainant du polymorphisme d'insertion entre les clones étudiés a été établie. Elle permet d'avoir une idée du polymorphisme généré pour chacun des éléments transposables et donc de mieux les sélectionner pour étudier la diversité de la vigne. Par exemple, *Vine-1* n'a pas été identifié comme actif dans notre étude. Il fut également utilisé par Verries *et al*. (2000) et aucun polymorphisme entre les clones étudiés n'avait été révélé. Nous avons choisi pour nos études portant sur la diversité des clones, d'utiliser des éléments transposables avec un haut niveau de polymorphisme d'insertion détecté entre les clones séquencés comme *Gret1, Copia-10, Gypsy-19, Caul-1*. Une étude pourrait aussi être réalisée à l'aide d'éléments ayant un polymorphisme d'insertion intermédiaire.

6.3-Etude préliminaire de la diversité du Pinot à l'aide des éléments transposables

La diversité des clones de Pinot agréés a été étudiée à l'aide de quatre éléments transposables. La méthode S-SAP a été utilisée afin d'identifier les insertions de chacun des éléments transposables dans une partie du génome des clones. La S-SAP fonctionne comme un marqueur dominant (absence ou présence de l'élément) et ne permet donc pas la distinction des hétérozygotes. Nous avons cependant utilisé cette méthode car elle permet très d'avoir connaissance rapidement des différentes insertions d'un élément transposable dans le génome. Les quatre éléments transposables ont fourni un profil génétique unique pour chacun des clones de Pinot étudiés. Les premiers clones de Pinot datent probablement du Moyen Age. Ces quatre éléments transposables produisant un profil d'insertion unique par clone ont donc eu une activité d'insertion récente. Nous avons analysé les arbres phylogénétiques en considérant que les éléments transposables s'accumulent de façon stable au cours de la vie de la plante et des cycles de multiplication végétative. Cependant, des données obtenues dans cette thèse ainsi que des références bibliographiques montrent des réversions de certaines insertions (Kobayashi *et al.*, 2004; Fernandez *et al.*, 2007) ce qui pourrait provoquer alors un biais dans nos arbres phylogénétiques.

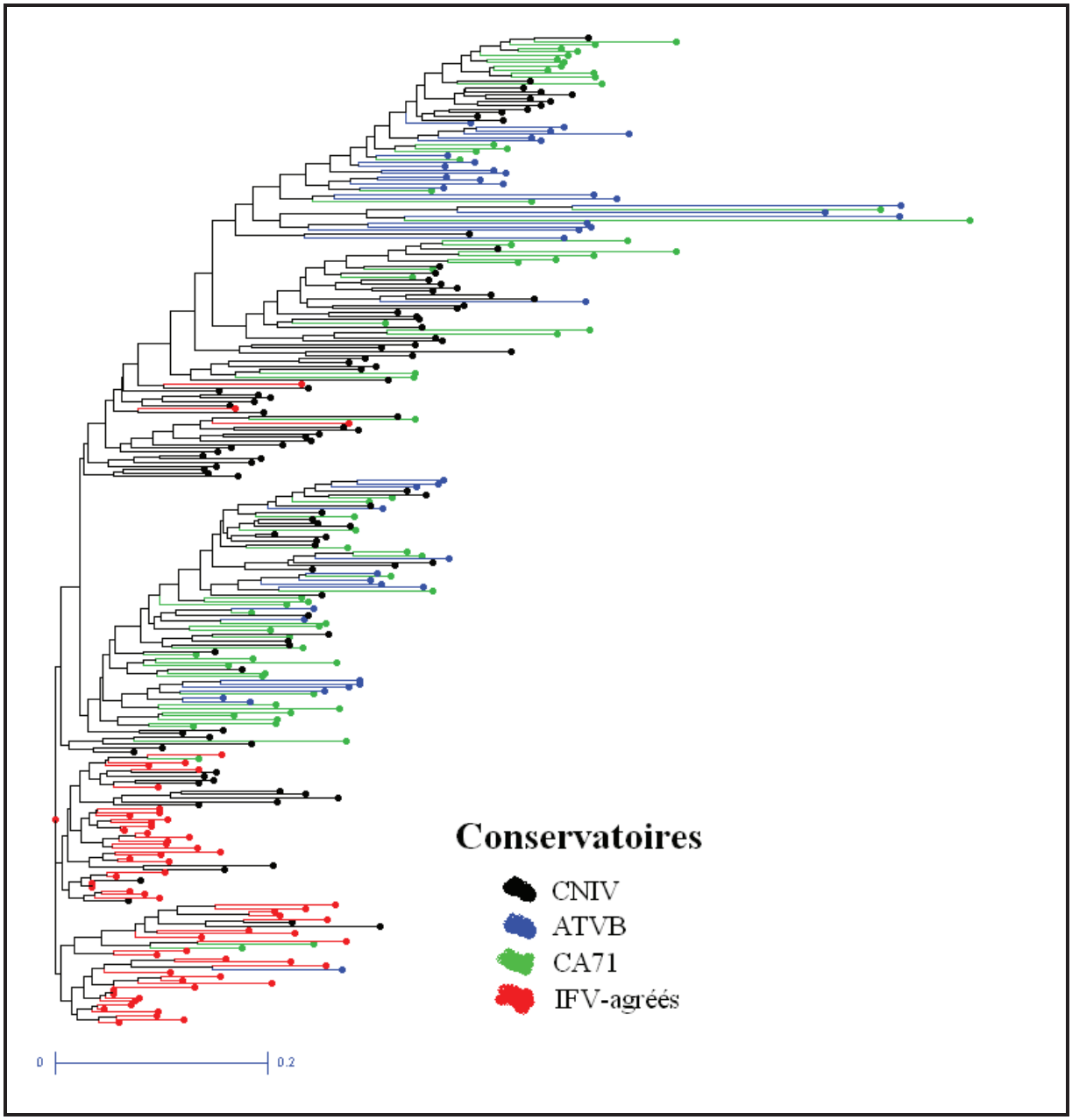


Figure 43 : Comparaison de la diversité des clones de Pinot présents dans les conservatoires et des clones agréés révélée par S-SAP de l'élément *Gret1.*

La structuration de la diversité des clones de Pinot obtenue à partir des profils S-SAP a été mise en relation avec les données géographiques disponibles. L'absence globale de structuration indique que des échanges de matériel végétal entre les différents bassins ont probablement eu lieu. Cependant les données géographiques disponibles ne sont pas suffisamment précises, ce qui limite la portée des conclusions.

La mise en relation de la structuration de la diversité des clones de Pinot par S-SAP et certains caractères phénotypiques qualitatifs (couleur, villosité), montre que le caractère meunier est probablement apparu à plusieurs reprises indépendamment dans différentes lignées de clones. Nous ne disposons actuellement que de peu de données phénotypiques. Une étude ultérieure effectuée à partir de données quantitatives (comme le rendement ou la qualité), et si possible répétée pour ne pas tenir compte de l'interaction avec l'environnement, permettrait d'effectuer une analyse plus rigoureuse.

Pour finir, la diversité des clones Pinot agréés a été comparée à celle présente actuellement dans les conservatoires de Pinot. La diversité dans ces conservatoires est bien supérieure à celle des clones agréés suggérant qu'il existe encore de nombreux clones dont la diversité génétique reste inexploitée. Depuis la prospection à l'agrégation il se déroule au moins 15 ans, le dernier Pinot noir agréés datent de 1989. Les clones agréés sont donc issus d'une première vague de sélection, qui aura cherché à valoriser des caractères d'intérêt très spécifiques limitant ainsi leur diversité.

Chapitre 4, Comparaison du polymorphisme clonal entre cépages

1-Introduction

Dans le précédent chapitre nous avons mis en évidence que l'activité des éléments transposables est la source principale de polymorphisme moléculaire à l'origine de la variation clonale. Cependant ces résultats ont pu être observés exclusivement chez le cépage Pinot et la vision du génome que l'on a pu étudier à l'aide du 454 GS-FLX reste limitée (approx. 1% du génome).

Grace à l'évolution des technologies de séquençage de nouvelle génération, nous avons pu étendre cette étude en étudiant de façon plus exhaustive un plus grand nombre d'individus. Onze clones de quatre cépages différents (Pinot, Syrah, Grenache et Sultanine) ont ainsi pu être séquencés et comparés. Nous présentons sous la forme d'un article en préparation les résultats de ces différentes comparaisons. Des résultats d'analyses supplémentaires seront ajoutés ultérieurement pour la publication. Un panorama quasi exhaustif des polymorphismes moléculaires qui s'accumulent durant la vie de la plante et au cours des cycles de multiplication végétative a ainsi été dressé. Le rôle de ces polymorphismes dans la diversité clonale et variétale a été évalué.

Les résultats de cette étude confirment entre autre, le rôle majeur de l'activité des éléments transposables dans la diversité clonale. Afin de mieux appréhender l'activité d'insertion de ces éléments, nous avons effectué, dans une perspective d'ouverture, deux études complémentaires au cours desquelles le polymorphisme d'insertion généré par quatre éléments transposables a été observé au sein d'un même individu et dans un échantillon de la diversité du genre *Vitis*. Ces études seront présentées à la fin de ce chapitre.

2-Présentation du matériel végétal

Dans cette seconde étude, la diversité clonale a été étudiée chez quatre cépages (Pinot, Syrah, Grenache et Sultanine). Ces cépages ont été choisis pour leur importance économique, puisqu'ils font partie des douze cépages les plus plantés dans le monde (J.M Boursiquot Com. Pers.). Ils ont également été choisis car ils n'ont pas de liens de parenté directs (Lacombe *et al.*, In prep) et enfin parce que nous disposions de clones particuliers pour chaque cépage présentant un caractère d'intérêt pour la sélection.

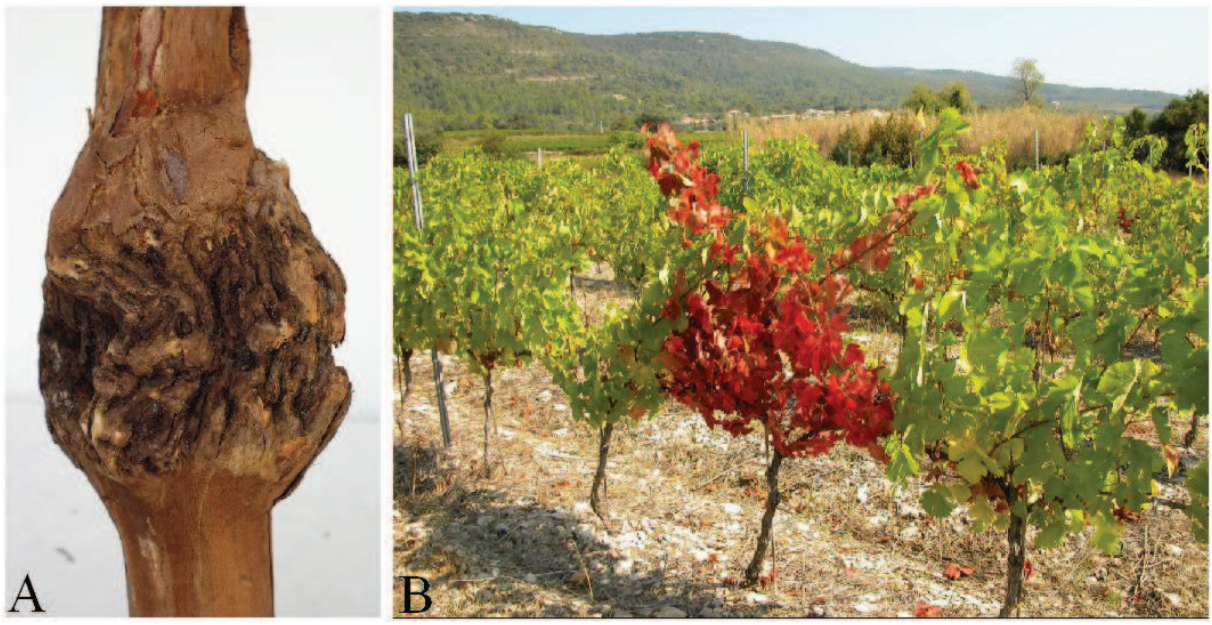


Figure 44 : Dépérissement de la Syrah. A) Nécrose au niveau du pied de la souche ; B) Symptôme de rougissement des feuilles (Photographie IFV).

Afin que les résultats obtenus dans nos deux études puissent être comparés, nous avons analysé les mêmes clones que précédemment: clones ENTAV-INRA® n°386, 583 et 777. Nous avons également ajouté deux clones de Pinot aux phénotypes d'intérêt. Le clone E52 a été choisi pour son port architectural particulièrement vertical. Ce caractère est recherché par les viticulteurs car il facilite les travaux agricoles. Le clone ST (Semi Teinturier) a été sélectionné pour son caractère lié à la couleur des baies, c'est à dire qu'une partie de la pulpe de la baie est colorée.

Le cépage Syrah a été sélectionné parce qu'il a une importance régionale forte et parce que l'IFV est préoccupé par un phénomène de mortalité grave affectant ce cépage appelé dépérissement de la Syrah. D'un point de vue physiologique, ce dépérissement se traduit par l'apparition de crevasses (nécroses) au niveau du point de greffe (Figure 44), suivi d'un rougissement foliaire entrainant généralement la mort de l'individu dans les deux à trois ans. Le dépérissement est observé dans la quasi-totalité des pays où on retrouve la Syrah (Spilmont, 2005) et des observations réalisées en Argentine en 2008 et au Chili en 2009 ont montré des phénomènes de dépérissement sur des Syrah plantées franc de pied (Spilmont *et al.*, 2010). L'hypothèse actuelle la plus probable du dépérissement de la Syrah est une origine génétique ou épigénétique. Certains clones se sont avérés beaucoup plus tolérants à ce phénomène, d'autres au contraire sont qualifiés de sensibles et ne sont plus commercialisés (Spilmont *et al.*, 2010). Un clone qualifié de très sensible le E266 et un clone tolérant ENTAV-INRA® n°470 ont été choisis pour notre étude. Les résultats du séquençage seront utilisés prochainement afin d'identifier d'éventuels polymorphismes corrélés au phénomène de mortalité.

Le cépage Grenache a été choisi car nous possédons dans la collection du Domaine de l'Espiguette (IFV), un clone au phénotype particulier. Ce clone appelé « Béru » possède des baies de petite taille, à l'opposé du clone n° 70 possédant des baies de taille normale. Une petite taille des baies est un caractère qualitatif recherché des viticulteurs. En effet, en général, plus la baie est petite, plus la qualité est importante (Champagnol, 1984). Les résultats obtenus pourront aussi être utilisés afin d'aider à la recherche des bases génétiques de ce caractère.

Enfin, le cépage Sultanine ne possède qu'un clone agréé (n° 919). Cependant nous disposons également de la variété Gora Chirine qui s'avère, d'après l'étude de son génotype (Laucou *et al*., 2010) être un clone de la Sultanine. Le Gora Chirine produit des baies au pH

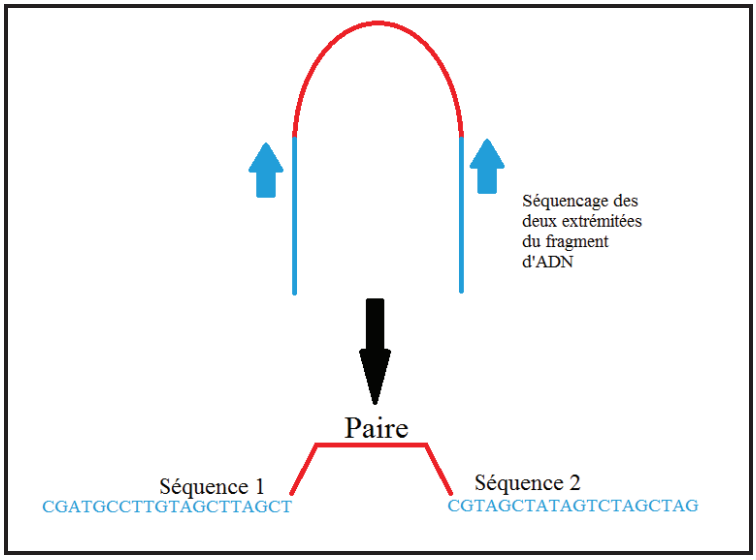


Figure 45 : Technologie "Paired-End".

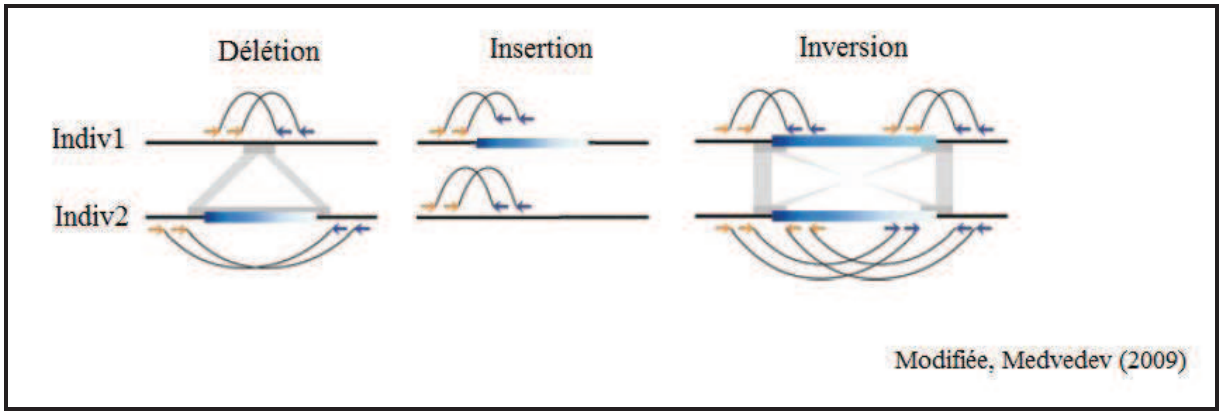


Figure 46 : Identification de variations structurales avec la technologie "Paired-End".

très faible, au contraire de la Sultanine n° 919 (Diakou *et al.*, 1997). L'acidité est un caractère qualitatif très important dont les bases génétiques chez la vigne sont encore peu connues (C. Romieu Com. Pers.). La comparaison de ces deux clones permettra de dresser une liste de gènes candidats jouant un rôle dans l'acidification de la baie.

Les données ainsi produites pourront donc être utilisées dans un deuxième temps afin de rechercher ultérieurement des gènes candidats corrélés aux caractères d'intérêt. Au total, les génomes de onze clones ont donc été étudiés (cinq clones de Pinot noir, et deux clones de Grenache, de Syrah et de Sultanine).

3-Choix de la méthodologie de séquençage nouvelle génération, 2010

En été 2010, lorsque nous avons envisagé les séquençages, le 454 était toujours en version Titanium (1,2 x 10^6 séquences de 400 bases) et l'Illumina était passé en version HiSeq 2000 (2 x 10^9 séquences avec une longueur de 100 bases en "Paired-End"). Le "Paired-End" consiste au séquençage des deux extrémités d'un court fragment d'ADN (200 - 400 bases). On obtient ainsi une paire constituée de deux séquences (Figure 45). La technologie "Paired-End" peut être utilisée afin d'identifier les variations structurales entre deux individus (Medvedev *et al.,* 2009 ; Figure 46).

Dans cette version, Illumina HiSeq 2000, deux cellules de séquençage (ou flowcell) comportent 8 lignes. Chaque ligne de séquençage permet d'obtenir environ 50 millions de colonies. Ainsi pour le génome de la vigne, une ligne de séquençage permet d'obtenir, en théorie, une couverture de 20X. Au vu du nombre de séquences produites et de l'augmentation de la taille des séquences, nous avons choisi de l'utiliser pour ces nouveaux séquençages au détriment du 454. Nous avons eu encore la chance d'être sélectionnés pour être projet pilote sur la plateforme GenoToul et les résultats de nos premiers séquençages ont été obtenus en février 2011.

4-Importance du polymorphisme moléculaire chez la vigne

Genome wide studies reveal high clonal and cultivar diversity in grape

Asexual and sexual polymorphism in grape

Carrier G.[1,2], Maillol V.[2], Pajeile M.[1], Ruiz M.[3], Romieu C.[2], Bouchez O.[5], Audeguin L.[1], Chatelet P. [1,2], Boursiquot J.M.[1,4], This P.[1,2], Le Cunff L.[1]

1. UMT Geno-Vigne®, IFV-INRA-Montpellier SupAgro, 2 place Viala, 34060 Montpellier, France
2. UMR AGAP INRA, F-34060, Montpellier, France
3. UMR AGAP , CIRAD, F-34398 Montpellier, France
4. UMR AGAP Montpellier SupAgro, F-34060, Montpellier, France
5. Plateforme Génomique de Toulouse Midi-Pyrénées, INRA Auzeville 31326 Castanet-Tolosan, France

Abstract

Grapevine clones are obtained through vegetative propagation from a single selected vine. In theory, vegetative propagation allows genome conservation. Nevertheless, after several propagation cycles, clones can acquire some traits differing from the original phenotype thus giving rise to clonal diversity. Clonal diversity in grape is used to select the best clones for commercial purpose as it is the only solution to access plant diversity without modifying the identity of cultivars with worldwide repute. Clonal polymorphism results mainly from accumulation of somatic mutations during plant growth. New generation sequencers open the way for a description of genomic mutations accumulation in a perennial plants.

The aim objective of this study was to provide a broad description of mutations accumulated by vines during their life and to evaluate their role in generating clonal and cultivar diversity. Eleven clones among four cultivars with high agronomic value were re-sequenced using a NGS approach. Data were analyzed with a new pipeline described in this study. Somatic polymorphism between two clones was composed with a low number of SNPs, while indels and structural variations displayed high polymorphism. We compared the number of polymorphism between two clones and between two cultivars and observed two molecular mechanisms. SNPs and indels accumulate continuously indeed structural variation this result suggests an elimination of this polymorphism. Molecular diversity accumulated along life of plant is a source of significant diversity and grape diversity could not be so high if these mutations do not exist.

Introduction

Vegetative propagation allows preservation of a given genome resulting from sexual reproduction. Some grape cultivars with worldwide reputations such as Pinot, Savagnin, Gouais and Cabernet franc have been multiplied vegetatively since the Middle Ages (Bowers *et al.*, 1999; Regner *et al.*, 2000; Boursiquot *et al.*, 2009) and their genomes have thus been preserved over several hundred years. However, after several vegetative propagation cycles, some individuals acquire distinctive traits giving rise to clonal diversity. This phenotypic polymorphism is used to select the best clones for commercial purposes and is of particular importance to the wine industry, which has been performing clonal selection programs in traditional cultivars for the last 45 years. Clone selection is the only way to improve planting material without modifying the identity of cultivars with worldwide reputation. The economic impact of clonal selection is therefore very high: for example approx. 95 % of the grapevines produced in French nurseries originate from the French clonal selection program. Clonal selection provided different clones such as Grenache ENTAV-INRA® n°70, Syrah ENTAV-INRA® n°470 and Pinot noir ENTAV-INRA® n°777 which all produce wines of superior quality (Boursiquot *et al.*, 2007).

Until now, clonal selection was mainly based on sanitary and phenotypic selection, and required a long and costly process (about 15 years) before having a good view of the real interest of the new clones. Similarly, clonal diversity, the driving force of selection, is difficult to assess *in-situ* and its analysis requires very large experimental designs to be pertinent. Comprehension of the basis of clonal variation and eventually identification of markers for molecular assisted selection (MAS) would thus be of high interest for the industry.

The main hypothesis to explain clonal diversity is the accumulation of mutations in somatic cells along life of plant and cycles of vegetative reproduction (Pelsy, 2009). SNP and short indels, currently the most studied polymorphism, could have potential impact on phenotypic variations (McCarroll *et al.*, 2008) in particular non synonymous SNPs (Ramensky *et al.*, 2002; McNally *et al.*, 2006). For example, one polymorphic SNP located in an exon generated a high production of linalol and geraniol responsible for muscat aroma (Emanuelli *et al.*, 2010). Transposable element activity, which drives genome evolution (Kazazian, 2004) also plays an important role in mutation and genomic reorganizations

(Kidwell, 2002). Several phenotypic variations known in grape clones such as the white color of skin berries (Kobayashi *et al.*, 2004) are the result of transposable element insertion. Another polymorphism could originate from epigenetic mechanism such as cytosine methylation which plays a role in the regulation of chromatin structure, mobile element activity (Mirouze *et al.*, 2009) and gene expression (Lauria & Rossi, 2011).

Grape clonal variants are also known to be of chimeric origin (Fernandez *et al.*, 2006; Moncada *et al.*, 2006; Walker *et al.*, 2006). Dicots have stratified shoot apical meristems which appear to be composed of two genetically distinct cell layers, two in grapes L1 and L2 (Einset & Pratt, 1954; Thompson & Olmo, 1963). Each cell layer remains developmentally independent from the adjacent layers and gives rise to different plant tissues: epidermis for L1 cells and internal tissues and gametes for L2 (Neilson Jones, 1969). A mutation occurring during plant development in one of the layer would produce a chimeric plant (Franks *et al.*, 2002). The chimeric nature of Meunier, a clone of Pinot noir displaying abnormal amount of hairs on the upper face of the leaves was demonstrated after separation of the two cell layers: plants regenerated from the L1 cell layer displayed dwarfism while plants regenerated from the L2 cell layer were phenotypically similar to Pinot noir (Boss & Thomas, 2002). Finally structural variation such as large deletion, insertion, inversion or duplication, (Hurles *et al.*, 2008), copy number variation, (CNV ; Freeman *et al.*, 2006; Stranger *et al.*, 2007), gene duplication (He & Zhang, 2005) and genomic disorder (Inoue & Lupski, 2002) could as well be at the origin of somatic mutations.

Previous studies of grapevine clonal diversity revealed several types of polymorphisms : SSR markers displayed limited clonal polymorphism (Hocquigny *et al.*, 2004; Moncada *et al.*, 2006). One study performed S-SAP approach with markers using universal retrotransposon-based primers and revealed polymorphism insertion (Wegscheider *et al.*, 2009). Until quite recently, no genowe-wide estimation of clonal polymorphism was available. A previous study (Carrier *et al.*, submitted) using 454 methodology proposed a comparison of 3 clones of Pinot noir and indicates that few SNPs and Indels were identified (a mean of 5.1 SNPs, 1.6 Indels per Mb) whereas much of the insertion polymorphism was generated by transposable element activity (35.1 mobile elements per Mb). In this study we provide a more exhaustive description of these somatic mutations. Polymorphism accumulation was identified and quantified to evaluate their importance in clonal and cultivar diversity. Eleven clones of four different cultivars were re-sequenced with NGS technology. To analyze the sequence data, we constructed “Bacchus pipeline” described in this study and downloadable at

marmadais.cirad.fr. This pipeline allows to genome reconstruction from reference genome by alignment method. Genome reconstruction quality evaluates and compares with reference genome. Finally, SNP and indel polymorphism between different individual is identified and used paired-end technology to identify structural variation such as described by Hajirasouliha *et al.* (2011) and Medvedev *et al.* (2009).

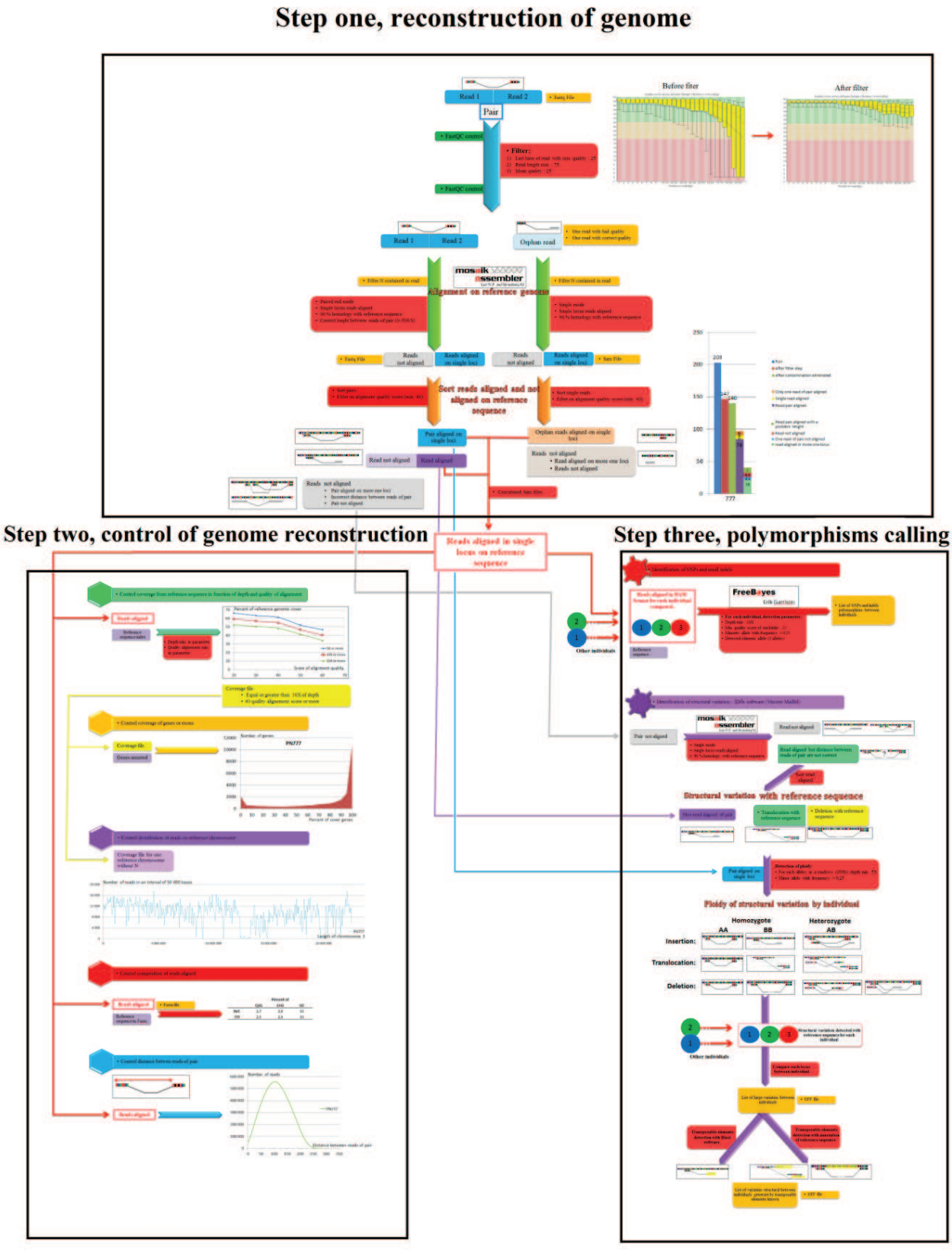


Figure 1: Global description of Bacchus pipeline for polymorphism identification from NGS (See Material and Methods).

Material and methods

Plant material and DNA extraction

Eleven clones were selected among four cultivars of *Vitis vinifera* L. (Pinot noir, Grenache, Syrah, Sultanine) to examine molecular diversity between clones of the same cultivar and between cultivars themselves. For Pinot noir, five clones were selected, based on their maximum phenotypic diversity: three certified clones ENTAV-INRA® n°386 (PN386), n°583 (PN583), n°777 (PN777) previously analyzed (Carrier et al, submitted), and two non-certified clones, one grown at the Domaine de l'Espiguette repository (PNE52), the other (PNst) grown at the Domaine de Vassal repository. For Grenache, two clones grown at the Domaine de l'Espiguette repository were selected: the certified clone ENTAV-INRA® n°70 (Gr70) and the non-certified clone "Beru". For Syrah, one certified and one non-certified clones grown at the Espiguette repository, were selected, respectively ENTAV-INRA® n°470 (S470) and E266 (SE266). For Sultanine, we used one certified and one non-certified clones grown at the SupAgro repository respectively ENTAV-INRA® n°919 (Sult) and Gora chirine (Gora).

Sequencing clone genomes

For each clone, we harvested 5 g of young leaves from a single plant. Nuclear DNA extraction was achieved using the NGS method previously described (Carrier *et al.*, 2011). Approximately 3-5 µg of nuclear DNA were used for Illumina HiSeq 2000 (Illumina corporation Inc.) sequencing as previously described (Linnarsson, 2010) at the Genotoul platform (INRA Toulouse). One lane was performed per individual. Reads were analyzed with FastQC software (v1.0) developed by Andrews S. in the Babraham Institute (www.bioinformatics.bbsrc.ac.uk) in order to validate run quality (sequence number, mean quality reads etc.). We obtained approximately 20 Gb (17 – 23 Gb) per run with a read length of 100 bases in paired-end and a mean sequencing quality score per read of Q37 (Ledergerber & Dessimoz, 2010). The average distance between the first nucleotide of the first and second reads was on average of 129 bases (supplementary figure 1). An average of 1.3 millions Illumina control reads and 3.2 millions reads from cytoplasmic DNA (3%) were present in each lane and considered in this study as contaminations.

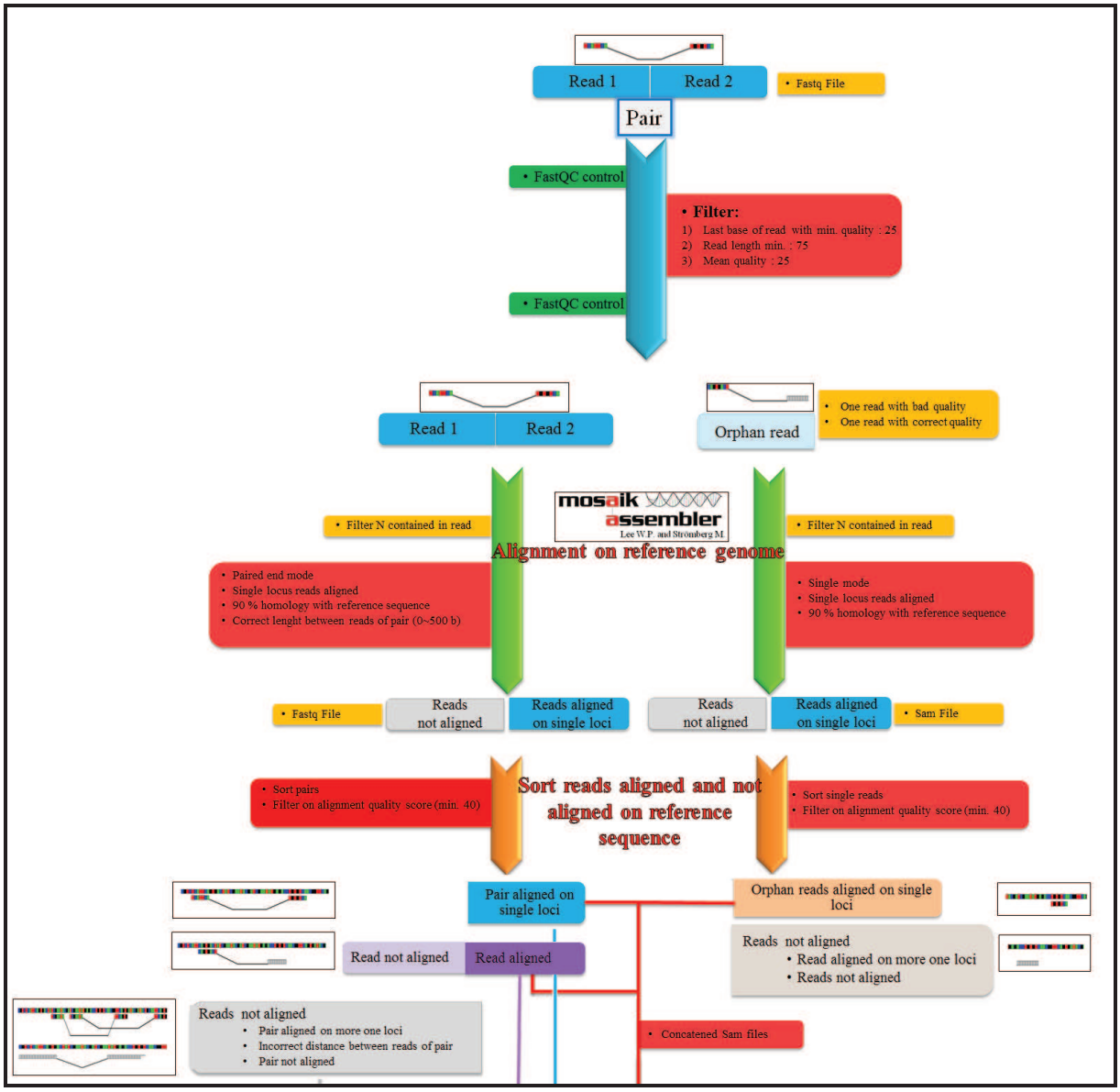


Figure 1-1: Description of Bacchus pipeline - Step one: reconstruction of genome.

Bacchus pipeline

This pipeline allows the identification of molecular polymorphisms between individuals sequenced with NGS. It is composed of three major steps: i) genome reconstruction, ii) control of genome reconstruction, iii) search for molecular polymorphisms. This pipeline uses the Galaxy interface (Blankenberg *et al.*, 2001). It can be used with data obtained with NGS using paired-end, mated-pairs or single method. With the paired-end method, more molecular polymorphisms are accessible than with the single method, in particular structural variations. We described in this study results obtained with Illumina HiSeq 2000 using paired-end method. This paired-end technology, with 100-base long reads, allows exploring all molecular polymorphism events except SSRs which have a length superior to 100 bases (For details on the Bacchus pipeline see Figure 1).

Genome reconstruction

Genome reconstruction was based on alignments on the grape reference genome used in its 12X version unmasked (12-Feb 2010, http://www.genoscope.cns.fr/). This step of the pipeline is composed of three parts: i) Filter on reads, ii) Read alignment on the reference sequence iii) Sorting of aligned and non-aligned reads on reference genome (Figure 1).

Reads obtained with NGS technology were filtered to retain only reads with sequencing quality over Q25 (Li, H *et al.*, 2008). First, reads are filtered by quality score (QS) of the last nucleotide of the read because quality decreased proportionally with read length (Dohm *et al.*, 2008; Carrier *et al.*, submitted). The last nucleotides which have a quality score lower than Q25 are eliminated until a nucleotide with a quality score of Q25 is detected. Reads are then filtered by length (reads smaller than 75 bp are eliminated) and in a third step, reads are filtered on mean quality score over all their nucleotides (QS > Q25). After this filtering step, we obtained for paired-end technology : i) pairs composed of two reads with acceptable quality and ii) orphan reads which provided only one read of acceptable quality in a given pair (Figure 1 and 2).

Genome was then rebuilt with hash alignment method implemented in Mosaik-Assembler software (Wan-Ping Lee and Michael Strömberg, http://bioinformatics.bc.edu/marthlab/). Reads were aligned on the grape reference sequence available at Genoscope (v12X 22-jav 2010; http://www.genoscope.cns.fr/spip/). For alignment we used parameters recommended by Mosaik comceptors i.e. reads were aligned at single locus with 90% homology with

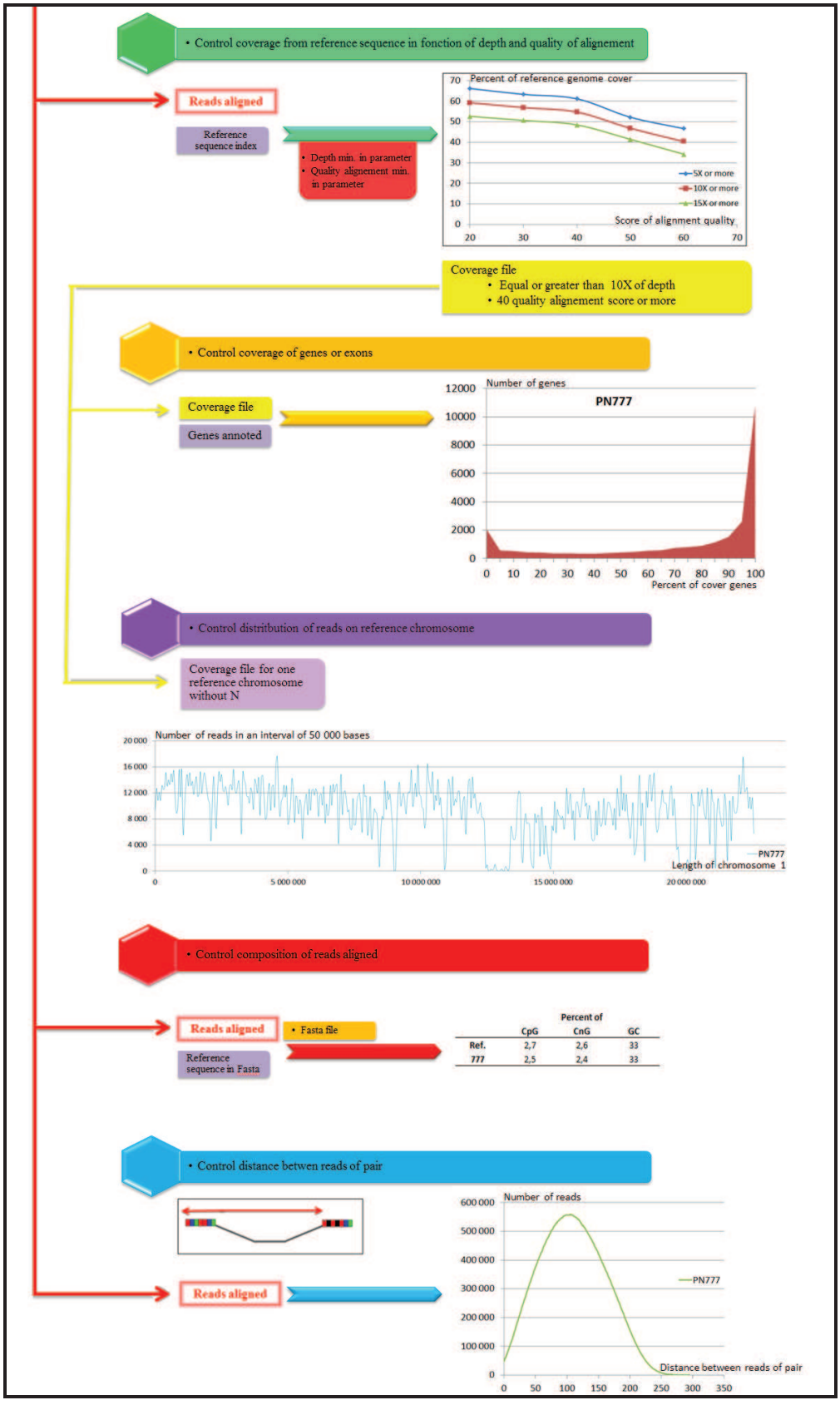


Figure 1-2: Description of Bacchus pipeline - Step two: control of genome reconstruction.

reference sequence. For paired-end reads, length between the two paired reads was checked by Mosaik-Assembler (length between reads of pair must be less than 500 bases). In parallel, orphan reads were aligned with single method parameters (Figure 1).

Aligned and non-aligned reads were sorted in different categories. For paired-end reads: i) pair with both reads aligned ; ii) pair with only one read aligned and iii) paired not aligned because both reads could align on more than one locus or length between both reads of pairs was too long (>500 bases). For orphan reads, only two categories were distinguished: reads aligned on a single locus and non-aligned reads. Only reads aligned on a single locus with an alignment quality score of 40 or more were retained as correctly aligned reads (Li, H *et al.*, 2008). Finally, all aligned reads (fully aligned pairs, one read of pair aligned and orphan reads aligned) were pooled to be checked and to identify polymorphisms.

Control of genome reconstruction

Aligned reads were compared with the reference sequence in five steps: i) control of the reference genome coverage according to quality score alignment and depth coverage. To be qualified as covered, one nucleotide must have a coverage rate equal or greater than 10X of depth and an alignment quality equal or greater than Q40 (Harismendy *et al.*, 2009). ii) Estimated of the annotated gene coverage in the reference genome which must be equal to or greater than 10X of depth. iii) Control of the distribution of reads on a reference chromosome. iv) Comparison of the composition of the aligned sequences (Percent of GC, CpG, CnG). v) Control of the distance between each read of a pair.

Identification of polymorphism

We identified molecular polymorphisms at three levels: within each individual, between pairs of clones from single cultivar and between cultivars. Since grapevine is highly heterozygous, we first searched all heterozygous SNPs and indel sites within one individual. Identification of polymorphisms between clones was then performed by comparing all pairs of reassembled sequences from a single cultivar (13 pairs). Identification of polymorphisms between cultivar was the performed selecting the best sequence from each cultivar (6 pairs).

SNPs and short indels

Reads aligned of each individual were compared to identify SNPs and indels using FreeBayes software (Garrison E., available at bioinformatics.bc.edu/marthlab/) which also

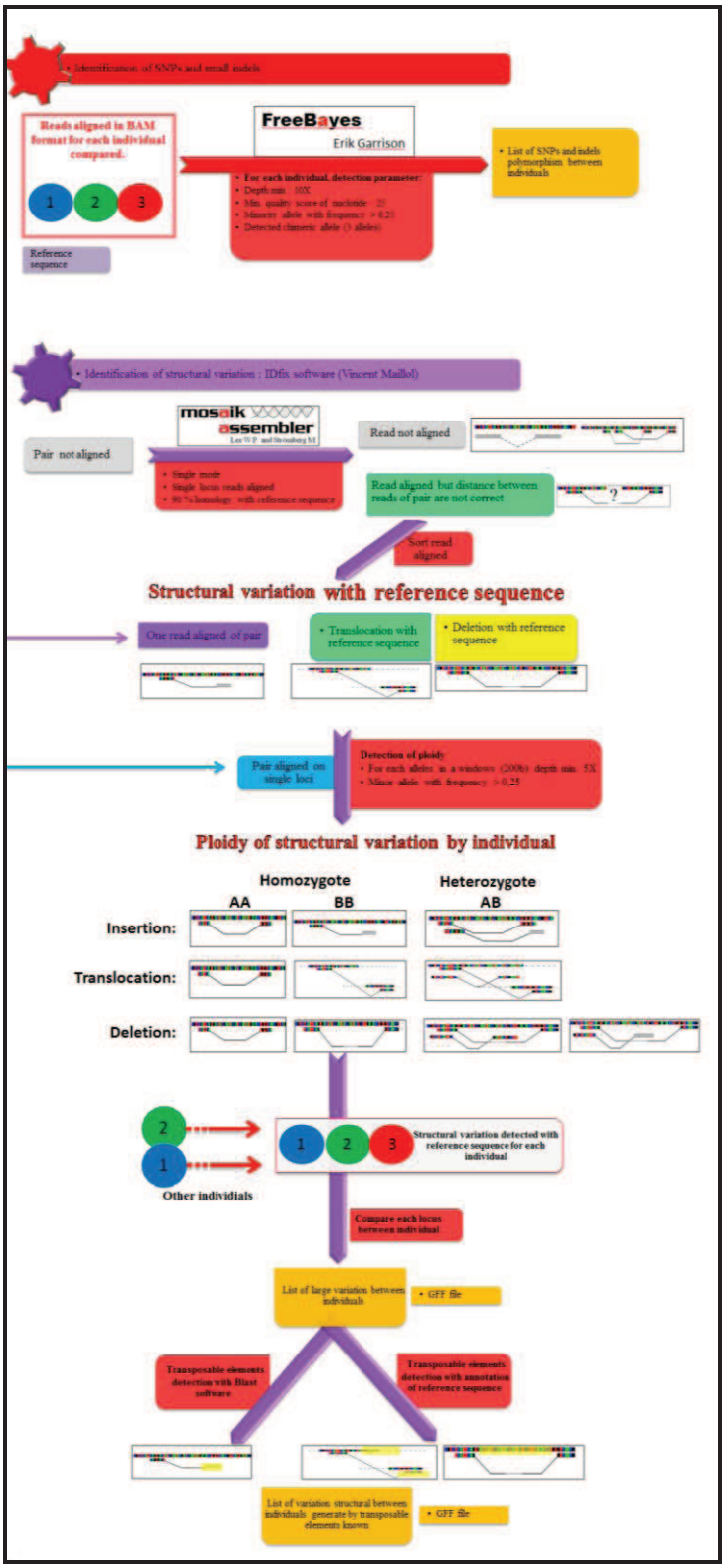


Figure 1-3: Description of Bacchus pipeline – Step three: Polymorphism calling.

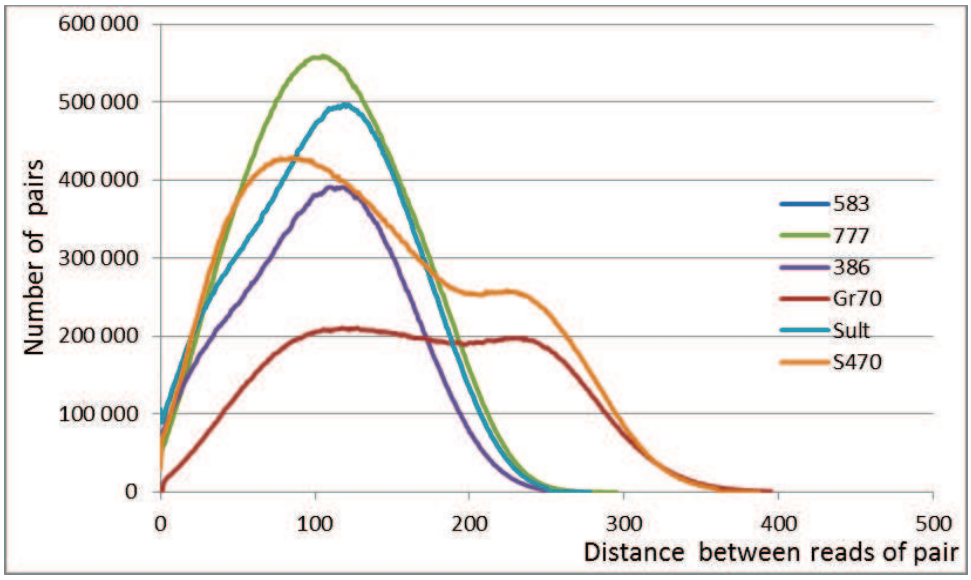
allows detecting SNP or indel polymorphism with three alleles. We set the threshold of the analysis: to detect polymorphisms, common regions must have a read depth equal or higher than 10X and a nucleotide quality score equal to or greater than 25. Moreover all minor alleles detected must have a depth frequency greater than 0.25 for each individual. These conditions reduced false positives (Harismendy *et al.*, 2009; Hedges *et al.*, 2009; Benaglio & Rivolta, 2010). To be considered, indel polymorphisms must localize within reads and have a length shorter than 20 bases.

Structural variations

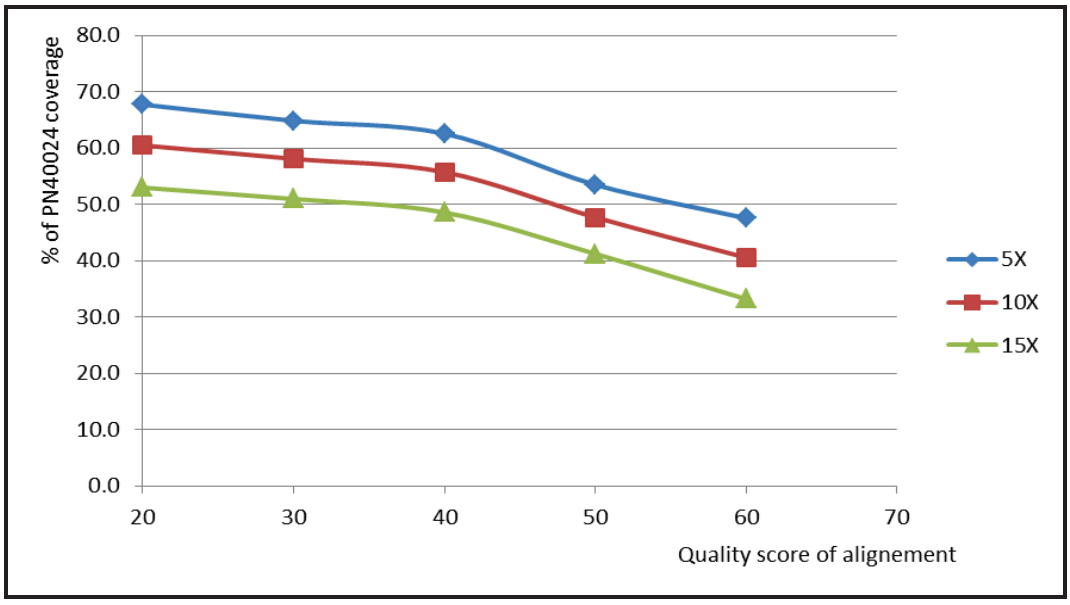
Paired-end method was used to identify structural variations. We inspired of (Hajirasouliha *et al.*, 2011) and (Medvedev *et al.*, 2009) works to detect structural variation. During genome reconstruction, some of the pairs could not be aligned because the distance between both reads was too large (deletion compared with the reference sequence) or because reads in a given pair were not located on the same chromosome (translocation in comparison to reference ; Figure 1). First, non-aligned pairs were considered like single reads and re-aligned in single mode. We then selected pairs which had two reads aligned on a single locus on the reference genome and identified each locus displaying structural variation. To detect other structural variations in comparison to the reference sequence, pairs with only one read aligned were used. These data allow the detection of each extremity of a structural variation. For each structural variation detected with reference sequence, only loci composed of five or more pairs with structural variation bases were retained. A heterozygous locus with a structural variation in a given individual was considered as such if minority alleles had a depth frequency higher than 0.25. Finally, each locus containing structural variation was compared between two individuals sequenced to detect structural variation polymorphisms themselves.

Detection of transposable elements

Structural variations such as insertion or deletion may be generated by transposable elements activity. For structural variation detected with pairs composed of one aligned and one non-aligned read, we searched whether the non-aligned sequence had a homology with a known grape transposable element. Blast software (Altschul *et al.*, 1997) was then used to identify sequence homologies between the non-aligned reads and the grape transposable element sequences available in Repbase database (Jurka *et al.*, 2005 v.2011). For the other



Supplementary Figure 1: Distance between the first nucleotide of the first and second reads of aligned pair.

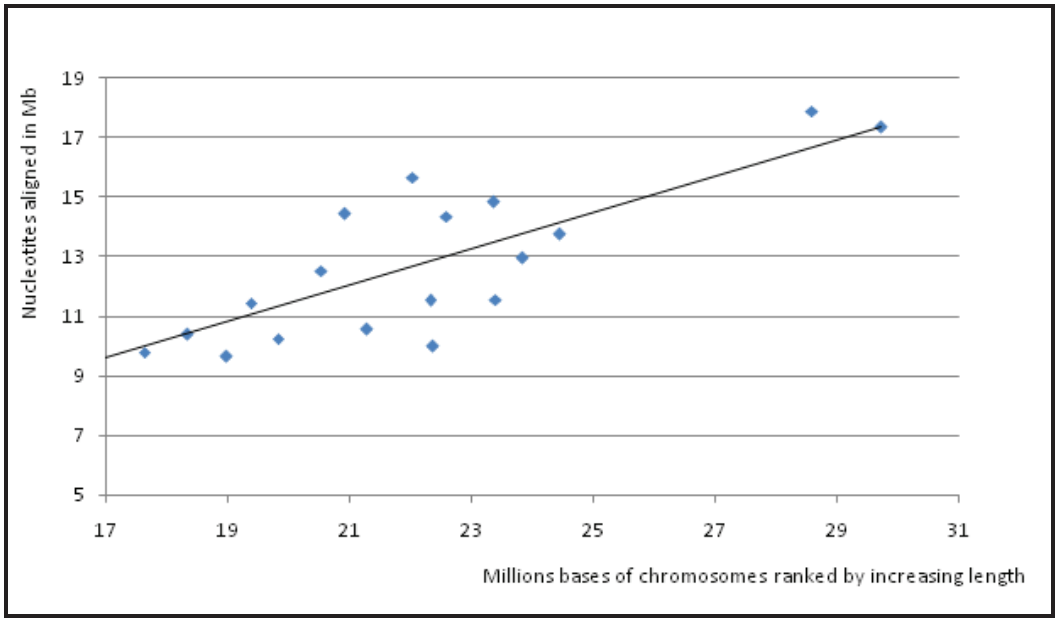


Supplementary Figure 2: Coverage of PN40024 by reads from PN583 in function of the depth of coverage (5X ; 10X ; 15X) and quality score of the alignment.

structural variations detected (deletion or translocation), we searched whether these variations were generated by known transposable elements and were annotated on the reference sequence (Figure 1). Annotation of known transposable elements on the reference sequence were performed using grape Repbase with RepeatMasker software (Smit *et al.*, 1996-2004).

	%CpG	%CnG	%GC
PN40024	2,7	2,6	33
777	2,5	2,4	33
386	2,6	2,5	33
583	2,8	2,8	34
PNst	2,4	2,4	32
E52	2,4	2,4	32
Gora	2,4	2,3	32
Sult	2,4	2,4	32
Gr70	2,4	2,3	32
Beru	2,4	2,3	31
S470	2,4	2,4	32
SE226	2,4	2,3	32

Supplementary Table 1: Composition of sequenced regions in percent of specific bases obtained with Illumina HiSeq 2000.



Supplementary Figure 3: Validation of random distribution of aligned reads. The estimated random correlation between the number of aligned reads and the length of the chromosome was tested using Pearson's correlation (R^2= 0.79, P-value < 0.05).

Results

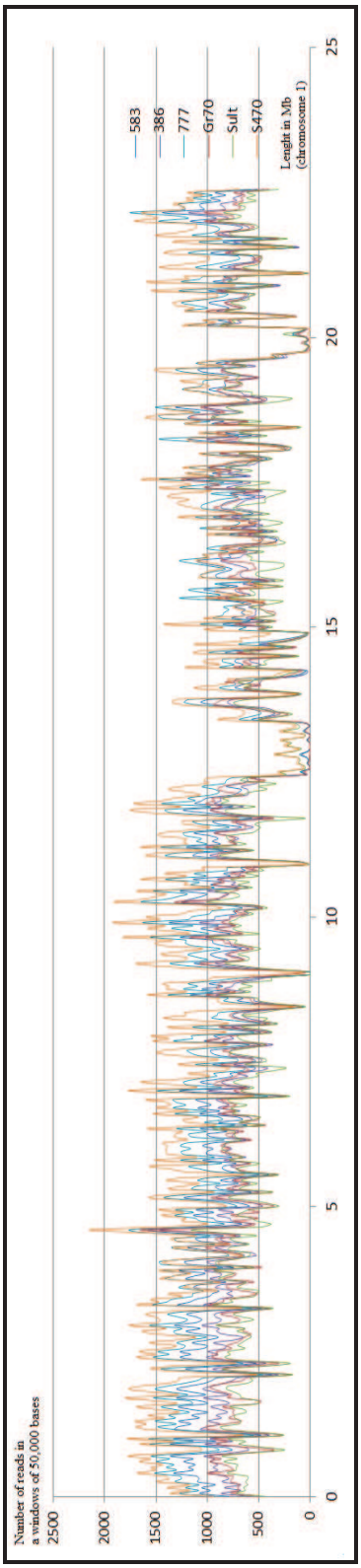
Alignment and comparison of sequenced genomes with the reference genome (PN40024)

We analyzed sequences obtained with Illumina HiSeq 2000 (v2.0) on the eleven clones. Averages of 191 millions reads (standard deviation (σ) =22) per lane were obtained (Figure 2). Reads were filtered on sequence quality score. Reads corresponding to contamination (8 millions (σ=1) in average) were eliminated. For each individual, averages of 116 millions reads (σ=21) were used for genome reconstruction (Figure 2) through alignment on PN40024 sequence. Genome reconstructions were performed with the Bacchus pipeline described in the Materiel and Method section.

From these 61% of the reads on average (75 millions (σ=12)) aligned on a single locus (50% pairs fully aligned; 5% with one aligned read per pair and 6% aligned orphan reads) representing a mean coverage of 52% of PN40024 (supplementary Figure 2). On average, 116 Mb of genic regions annotated in PN40024 (Jaillon *et al.*, 2007) were covered which represent 68% of genic regions. Similarly, a mean of 22 Mb (74%) of exonic regions, were covered.

Among unaligned sequences, some pairs or orphan reads (22%) were not retained due to their multiple putative localizations on the reference sequence. Some other pairs were not aligned (6%) during the genome reconstruction because the distance between both reads in a pair was too large or because the two reads did not align on the same chromosome. Furthermore, 6% of reads only did not match any known sequences of PN40024 (Figure 2). These sequences may correspond to absent regions of PN40024 or to contaminations not search during the analysis (viruses, fungus).

We compared several criteria (GC, CpG and CnG in the aligned sequences) between sequences obtained for the clones and for PN40024, no difference was observed (Supplementary Table 1). The number of bases aligned on each chromosome was proportional to their length (R^2 >0.79, P-value <0.05; Supplementary Figure 3). However, our results indicated that read distribution along the chromosomes was not random and some regions were consistently excluded from alignment (Supplementary Figure 4 for an example on chromosome 1). There was a significant negative correlation between the number of aligned sequences and the number of repeated elements in the sequences (correlation coefficient = -0.74 and p-value <0.01).



Supplementary Figure 4: Distribution of reads for 6 clones aligned on chromosome 1 obtained with Illumina HiSeq 2000. The uncovered region around 13 Mb in chromosome 1 corresponds to the centromere.

Polymorphism calling

Between haplotypes of each individual

Polymorphisms between the haplotypes of each individual were identified from aligned sequences. On average, polymorphism between haplotypes of a single individual (rate of heterozygosity) was estimated at 6 366 SNPs per Mb (σ=133) and 2 230 indels per Mb (σ=94). There was no difference in heterozygosity between the 4 cultivars. Moreover, triallelic loci were also observed and all the analyzed clones displayed on average 1.1 % and 8% of triallelic SNPs and indels respectively.

Between clones and between cultivars

Regions shared by all sequenced individuals represented on average 47% of the reference genome for a depth equal to or greater than 10X. Different genetic polymorphisms (SNP, indel, structural variation) were identified between clones of each cultivar (five clones compared for Pinot noir and two for the Grenache, Syrah and Sultanine cultivars) and between these four cultivars themselves (see Material and methods). The mean number of polymorphisms detected between two clones of the same cultivar were not different according to the cultivars (Confidence interval = 0.95). On average, 2.5 SNPs (σ=4), 11.5 indels (σ=1.8), 129.0 structural variations (σ=53) per Mb polymorphs were identified between any two clones (Figure 3). Polymorphisms detected between two cultivars were on average per Mb: 90 SNPs (σ=8.7), 66.4 indels (σ=5.2) and 176.0 (σ=67) structural variations (Figure 3). Among identified structural variations, 34% were generated by transposable elements already known in grape. Polymorphic transposable elements were classified as either class I (69%) or class II (31%) transposable elements. Other structural polymorphism composed of transposable element not identified, duplication or inversion.

Among these polymorphisms, the ratio of polymorphisms localized within genes or exons were not different between two clones of the same cultivar or between two cultivars (Confidence interval = 0.95). On average over the 11 clones, 22% and 1.5% of SNPs localized within genes and exons respectively, while 29% and 0.3% of indels, and 30% and 4% of structural variations located in genes and exons respectively (Figure 4). Globally, it means in average between any two clones, 6 SNPs, 3 indels and 1917 structural variations in the analyzed coding regions.

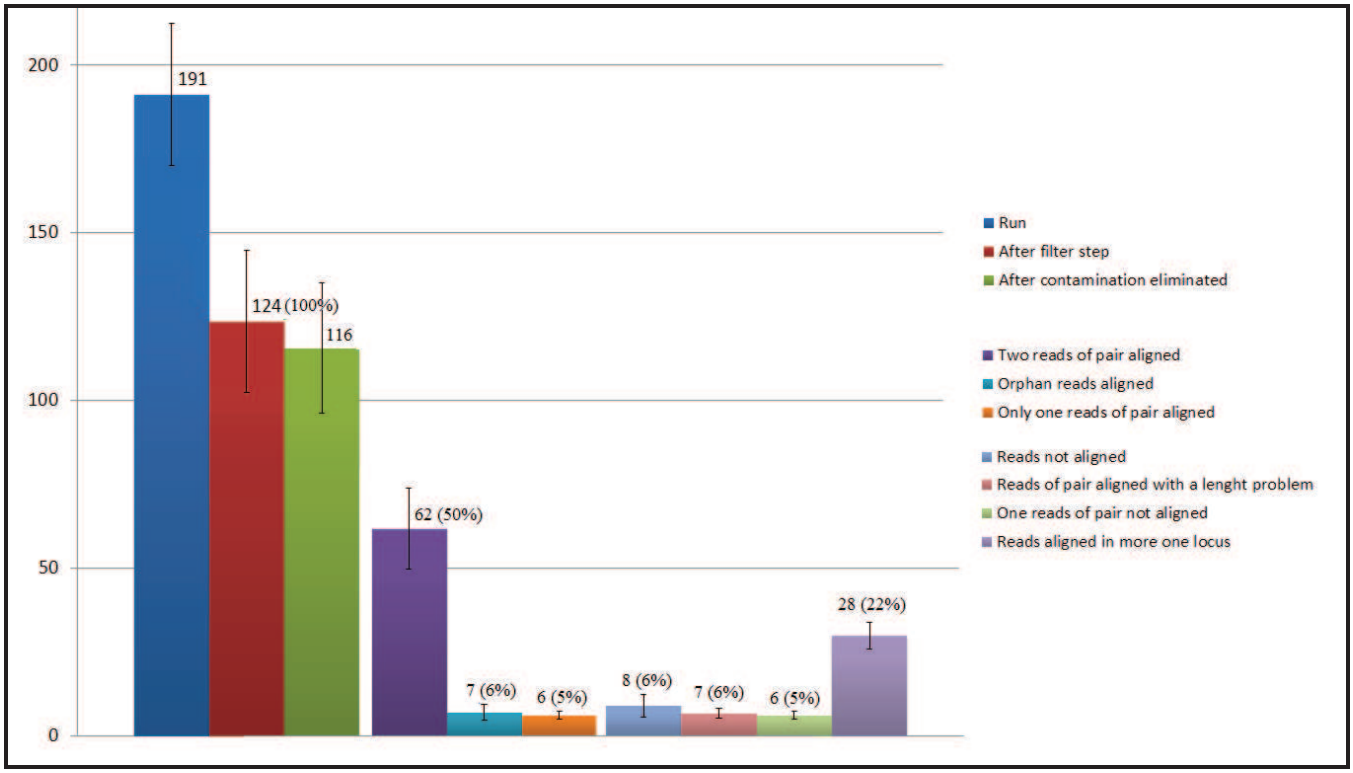


Figure 2: Average of reads obtained after the different filtering steps of Bacchus pipeline.

Comparison of clonal and varietal polymorphism

The polymorphism between two clones of the same cultivar (clonal diversity) was lower than the polymorphism between two clones of different cultivars (cultivar diversity) for all types of polymorphisms studied (Figure 3). A bigger difference in polymorphism rates between clone diversity versus cultivar diversity was observed for SNPs and indel (x36 and x7 respectively) than for large structural variations (x1.3). In the coding genome, the same pattern is observed with even amplified difference (x61, x13 and x1.3 SNPs, indel and structural variants respectively; Figure 4)

Discussions

In this study we selected four cultivars (Pinot noir, Grenache, Syrah and Sultanine) with no parentage relationships and with a high agronomic value (Lacombe *et al.*, In prep). All four cultivars are relatively old cultivars, even if Pinot is believed to be older. Polymorphism accumulation was identified and quantified to evaluate their importance in clonal and cultivar diversity.

Comparing sequenced genomes

Genome reconstruction

With a relatively straightforward method of alignment of the reads onto the reference genome, on average, 52% of the reference genome was covered and there was no significant difference of coverage between clones or cultivars. Our results are similar to the alignment obtained by Myles *et al*. (2010). Grape genome is composed at 41.4% of repeated regions (Jaillon *et al.*, 2007), making it impossible to obtain a coverage greater than 60% of the reference genome with this alignment method. Repeated region reconstructions such as centromere, telomere, and satellite regions by alignment method were thus difficult. These regions call for a *de-novo* approach for their reconstruction but this method requires significant computational power and is delicate with 100 base-long reads (Flicek & Birney, 2009). In this study we did not search SNPs or indels in repeated regions but used paired-end technology to detect structural variation polymorphisms generated by repeated sequences such as transposable elements.

One has however to keep in mind that the reference genome is a nearly homozygous sequence and that large structural variation between alleles, even within Pinot would not be mapped as well. Bacchus pipeline is presently under evolution, in particular in order to

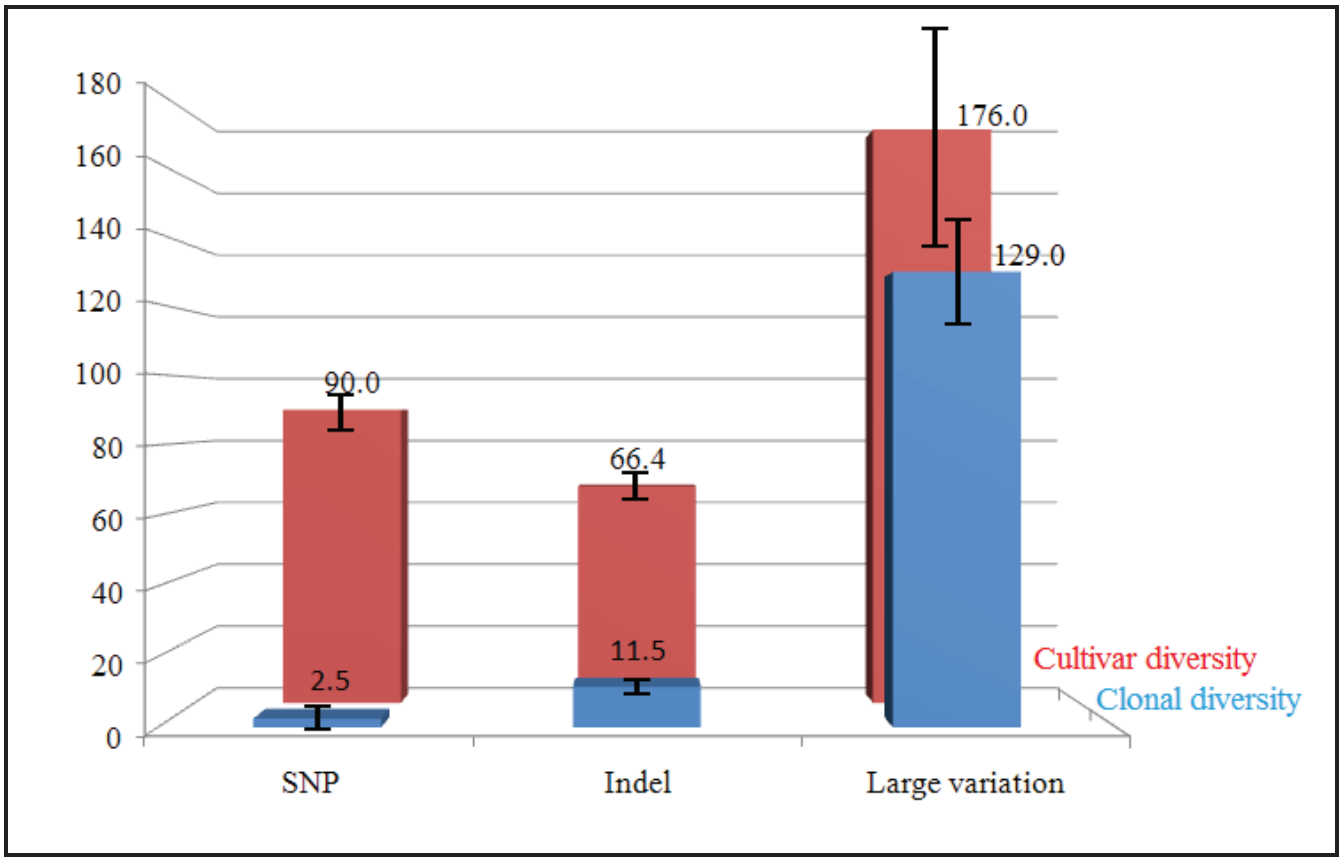


Figure 3: Average polymorphisms identified per Mb between clones or cultivars for the 3 classes of mutational events, SNPs, Indels and structural variants. For clones, the value are the means of the 16 pairs comparisons (10 for Pinots and 2 for the 3 others cultivars). For cultivars, the values are the means of 6 pairs comparisons.

develop *de-novo* approaches. Such approach will help us map more reads but also to search for other types of mutations such as CNV.

Polymorphism between haplotypes versus polymorphism between cultivars

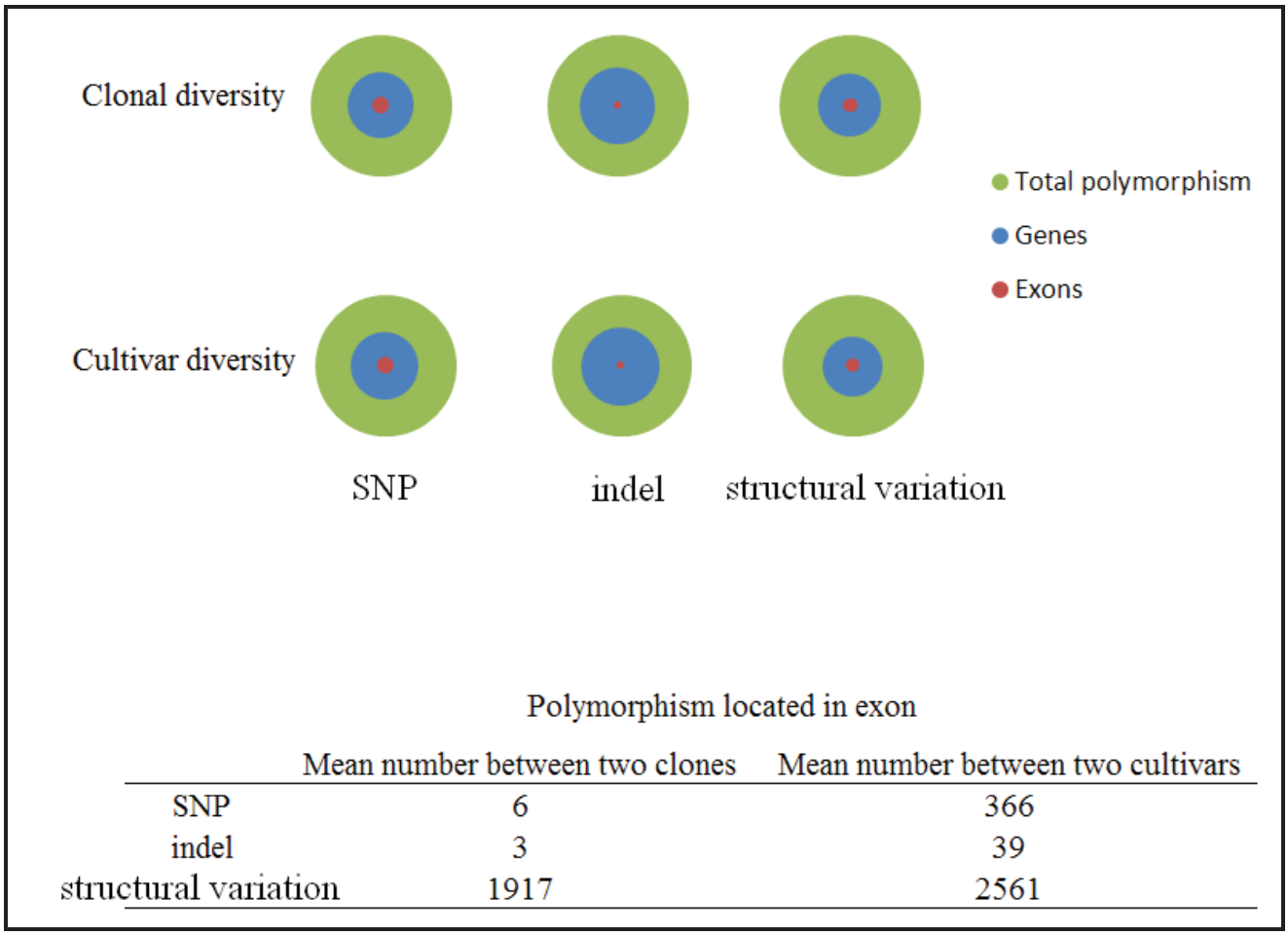
Polymorphism in haplotypes of each individual was estimated on average at 6 366 SNPs and 2 230 indels per Mb but with an analysis of 50% of the genome. This result is slightly higher than polymorphism measured in the Pinot noir ENTA-INRA n°115 genome by Velasco *et al*. (2007) (4255 SNPs, 2127 indel per Mb). A bias in our study could have been introduced because we only analyzed unrepeated regions. An inferior depth of coverage of PN115 sequence (6.4x on average) can also be at the origin of this difference. Benaglio & Rivolta, (2010) estimated that 15X of depth is necessary to identify 90% of the polymorphisms contained in the genome. The heterozygosity level was identical for the 4 cultivars which is in agreement with previous observation using 20 SSR markers by Laucou et al. (2011): a level of heterozygosity of 0.75 to 0.95 was estimated for these 4 cultivars. Our analysis thus confirms at a genome-wide scale the high level of heterozygosity of the grape genome. It is also surprising that polymorphism between both haplotypes of a single individual is almost 70 times higher than the polymorphism between cultivars (6336 versus 90 SNPs per Mb). Comparison of position of the heterozygous loci between cultivars should be performed to understand the significance of these differences.

Comparing sequenced genomes

The common sequenced portion of the genomes of the 11 clones, when assembled and compared, covered 47 % of the grape genome. It thus means that the sequencing method is not fully random and that several regions common to each clone was not represented in the sequence. Low-alignment regions showed over-representation of repeat elements in some areas, particularly at the putative centromere location.

Molecular polymorphisms accumulated somatically during plant life

A previous study described molecular polymorphisms in 3 Pinot noir clones (Carrier e*t al.,* submitted) considering a smaller fraction of the genome (~1%). Few SNPs and indels were detected and the most significant polymorphic elements were transposable elements. In this study in four different cultivars, we confirm that somatic polymorphism for SNPs and indels was low compared to structural variations. Moreover we identified structural variants



Polymorphism located in exon

	Mean number between two clones	Mean number between two cultivars
SNP	6	366
indel	3	39
structural variation	1917	2561

Figure 4: Distribution of inter and intra cultivars polymorphisms according to the type of mutations and their location in the genome.

without *a priori* and not only those generated by transposable elements already known in grape.

SNP polymorphims

SNPs in somatic reproduction results of error accumulations during DNA replication (Longley *et al.*, 2005) or repair (Fan *et al.*, 1999; Zienolddiny *et al.*, 2006). SNPs are accumulating in cells at a rate between 10^{-8} and 10^{-10} during mitosis (Baer *et al.*, 2007). In this analysis, we have identified more than 500 SNPs differences between clones in 47% of their genomes. By extrapolation around 1000 single mutation differences should differentiated clones. This is more than previously expected but in agreement with our first estimation (Carrier et al, submitted). Because SNPs are more permanent than other types of mutations accumulating in clones, and because they are sufficient to allow differentiation of clones, SNPs could be good candidates for clonal identification. More careful analysis of SNPs positions between clones could give us clues for rapid development of such markers in a large set of cultivars.

Indel polymorphism

Indels are also stable polymorphisms generated through DNA repair after a break caused by a crossing-over event (Dong *et al.*, 2002), insertion of an element (transposable element or other ; (Yamashita *et al.*, 1999; Van de Lagemaat *et al.*, 2005)) or a duplication (Chen *et al.*, 2005). In "vegetative" propagation, indels can only be generated by transposable element activity whereas in sexual reproduction indels are also generated by recombination events. SSR markers could also be classified in this type of mutations but only small SSR (50b) can be considered with this lenght reads (100b). However this type of marker have a length often larger than 100 bp, for this reason SSR were not consider in this first analysis. Polymorph indels is also quite numerous between clones since around 2500 indels in average were identified in 47% of their genomes.

Structural variations

As we had hypothesized the structural variants are the more frequent events differentiating clones. This genomic disorder is characterized by the presence of flanking segmental duplications that predispose these regions to recurrent rearrangements. Approximately 5% of the human genome is composed of segmental duplications (Sharp *et al.*, 2006). Grape is an ancient hexaploid genome and contain multiple duplicated regions (Jaillon

et al., 2007) among which 17% are segmental duplications (Giannuzzi *et al.*, 2010), suggesting that significant genomic disorder could be found in grape. The grape structural variation events studied in this work are only those polymorphisms generated by transposable element activity. Among structural variants identified in this study, 34% were generated by the activity of mobile element known in grape (38.7 insertions per Mb), in accordance with previous results (Carrier *et al*., submitted). The database of transposable elements known in grape is not yet complete and depending on addition of new elements, the proportion of transposable elements in structural variation could correspondingly increase. The number of such events made then quite easy to identify. Many transposable elements are however still active (Carrier et al, submitted) and additional data (data not shown) also demonstrated that they may be active even in a single plant generating polymorphism between tissues of the same plant. These markers may thus not be very useful for clonal identification.

Our data also underestimates the structural variants since CNVs and other large events such as duplication, invertions… have not been considered in this analysis. More complete analysis of the genomes is thus necessary.

Chimerism

Grape primordial are composed of two genetically distinct cell layers, L1 and L2 (Einset & Pratt, 1954; Thompson & Olmo, 1963). Each cell layer remains developmentally independent from the adjacent layers and gives rise to different plant tissues. Franks *et al.* (2002) have shown that when the two cell layers of the periclinal chimera are separated following somatic embryogenesis, the regenerated plants have DNA profiles distinct from those of the parental plant. But these chimerical mutations are not transferred to sexual progeny except if it considers L2 layer involved in gamete formation. By vegetative propagation, chimeric sector can be chosen to regenerate a new clone. In this study, all of the analyzed clones displayed a small percentage of tri allelic profiles. Such profiles are only observed when the mutation generates a third base and thus in most of the cases, chimeras are not identifiable. Nevertheless such polymorphism seems quite common in all analyzed clones.

Accumulation of somatic mutation does not differ between cultivars

Unexpectedly, we did not found any difference in the level of clonal polymorphism among cultivars. We have chosen these 4 cultivars because they corresponded to ancient ones. Pinot was however supposed to be much more ancient, since it is the ancestors of numerous contemporary varieties. Besides Pinot clones and mutants are much more numerous and

polymorphic in the collections than clones of any other cultivar. Other cultivar should thus be analyzed, in particular cultivar of recent origin such as recent selection of INRA (Marselan…) or progenies of Pinot or Sultanine in order to compare their polymorphism. If this is confirmed, then a important question should be asked: why is the level of polymorphism in a cultivar not correlated with the age of the cultivar. The phylloxera crisis of 19th century may also have erased longer accumulation of mutations and restarted the evolution clock of the clones.

If the number of mutation is not a pertinent factor, then the position of the mutations could also explain the differences in diversity observed in the field. More careful comparison of polymorphism site location should thus be performed.

Accumulation of somatic mutation is at the origin of clonal diversity

Clones are obtained by vegetative propagation from specific organ of a single selected vine. Throughout plant life, there is an accumulation of mutations in various tissues and organs and a plant is thus composed of a "mosaic of genomes" (Gill *et al.*, 1995). Vegetative multiplication thus allows selecting and propagating accumulated mutations contained in tissue or organ genomes. These mutations may have an impact on phenotype and generated a clone with few specific traits of particular agronomic interest. A small percentage of the mutations were observed in the genes and even fewer in the coding sequences (in average 27% in the genes and less that 2% in the coding sequences). In the sequences, genic regions correspond to roughly 33% and coding regions to 7%. There is thus a bias toward limitation of mutations in the coding sequences especially for indels (0.3 % only in coding sequences). Structural variants on the contrary were quite high (4%) but may also correspond to extinct genes. In the analysis, we however did not consider the presence of polymorphisms in promoter regions of the genes, which could also have a strong effect on gene expression as demonstrated for *VvMybA1* (Kobayashi *et al.*, 2005) or *VvTFL1A* (Fernandez *et al.*, 2010) gene. Such information should also be searched in the near future.

Nevertheless some of the identified polymorphisms in exons (Figure 4) could have an effect on phenotype of the clones. Few SNP were identified as a cause of phenotypic variation (Fournier-Level *et al.*, 2009; Emanuelli *et al.*, 2010). Location of identified SNP in exons will be searched more carefully in order to search for such effect. Similarly, a few indels are known to generate phenotypic polymorphisms (McIntyre *et al.*, 2008) but no such data is available in grape. However, indels are less frequently located in coding regions than SNP

because they more frequently generate deleterious mutations such as frame shifts (Ng *et al.*, 2008) leading to their elimination through natural selection.

The structural variation generated could finally also be at the origin of phenotypic variation (Inoue & Lupski, 2002; Hurles *et al.*, 2008). This variation has been poorly studied in plants but some studies in human research show that it could affect phenotype (Mefford *et al.*, 2007; Sharp *et al.*, 2007). In our study, 4% of structural variants are contained in genes which have a high probability to impact cell performance.

Some studies performed on causal mutations of phenotypic clones have showed the implication of transposable element insertions in the development of phenotypic traits such as white color in berries (Kobayashi *et al.*, 2005) or grape architecture (Fernandez *et al.*, 2010). But insertion of new copies of a transposable element is not the only event which generates polymorphism and can have an impact on phenotype. Elimination of a transposable element, known to be the consequence of non-homologous recombination, can also generate structural variation (Devos *et al.*, 2002; Moisy, C. *et al.*, 2008).

Moreover grape, composed of two genetically distinct cell layers, L1 and L2 (Einset & Pratt, 1954; Thompson & Olmo, 1963). Each cell layer remains developmentally independent from the adjacent layers and gives rise to different plant tissues. Franks *et al.* (2002) have shown that when the two cell layers of the periclinal chimera are separated following somatic embryogenesis, the regenerated plants not only have DNA profiles distinct from those of the parental plant but also have novel phenotypes such as the dwarf phenotype produced after regeneration from L1 cells in Meunier (Boss & Thomas, 2002). Chimerical mutations could generate sectorial phenotypic variations such as hairless sectors (Franks *et al.*, 2002) or, as in the case of Pinot gris, occasional production of colored berries with white variegations (Hocquigny *et al.*, 2004). But these chimerical mutations are not transferred to progeny except if this sector is chosen to regenerate a new clone. Accumulation of mutations and their selection by vegetative propagation is the origin of a large part of clonal diversity.

Difference of polymorphism generated by somatic and sexual reproduction.

The polymorphism observed between cultivars is much geater than the one observed between clones, but it is much pronounced for SNPs and indels than for structural variants. These results show that different molecular mechanisms are involved or at least that there is a mechanism regulating the larger events. SNPs and indels seem to have accumulated continuously between cultivars. On the contrary, our results suggest a quick reduction in structural variation polymorphism. Structural variation can entail major rearrangements

deleterious for cells and consequently be eliminated. Part of this structural variation is due to transposable element activity. Transposable element activity is very complex and varies depending on plant environment and stress level (Nagy & Chandler, 2004; Feschotte, 2008). Moreover, insertion generated by transposable element activities could be very quickly eliminated through complex recombinations (O'Hare & Rubin, 1983; Gray, 2000; Bowen & Jordan, 2002; Devos *et al.*, 2002). Some insertions of transposable elements are known to be hereditary. For example, insertion of *Gret-1* in the *VvMybA-1* promoter is known to be transferred to progeny (Fournier-Level *et al.*, 2010). However, the elimination of this insertion has also been observed in Chardonnay and lead to the presence of a solo LTR and rose berries (This *et al.*, 2007). Regulation mechanisms of others structural variations such as duplications, inversions, homologous recombinations are currently unknown in plants. Future study of this variation could open up a new way for polymorphism research.

Accumulation of somatic mutation impact in cultivar diversity

Genetic recombination arises from sexual reproduction. Cross-fertilization mutations increase genetic variability within the population, enhancing adaptation to a changing environment (Crow, 1992; Rice, 2002). Sexual breeding allows genome segregation between two individuals. It is at the origin of increased molecular and phenotype polymorphism compared to “somatic” breeding (Grapputo *et al.*, 2005). Indeed, all the polymorphisms identified were greater between cultivars than between clones. Considering the level of variation between haplotypes, their segregation in different cultivars could even have introduced more variation than observed here.

Myles *et al,* (2011) showed that grape has high phenotypic and nucleotide diversity although present varieties resulted from a few sexual reproduction cycles only. Nucleotide diversity of grape is evaluated at 0.0051 (Lijavetzky *et al.*, 2007) similar to maize (0.0063 (Ching *et al.*, 2002)), but higher than in other annual plants such as soybean (0.0012 (Zhu *et al.*, 2003)) or tomato (0.0010 (Labate *et al.*, 2009)). Studies by Klekowski *et al.* (1998) compared annual and perennial plants in the mangrove and concluded that progenies of perennial plants display a higher diversity than annual plants. Nucleotide diversity of grape cultivars is similar to that of annual plants whereas these latters have numerous sexual reproduction cycles compared to perennial plants.

Mutations accumulated throughout a perennial plant life, and in particular structural variations could be transferred to its progeny by sexual reproduction. But as gametes originate from L2 tissues, only mutation accumulated in this L2 layer could consequently be

transmitted to the progeny. Accumulation of mutations along plant life generates additional diversity which is amplified by the segregation of somatic genomes during sexual reproduction. This representation of perennial plant evolution has been modeled by Orive *et al*. (2001). Considering the level of polymorphism in clones, somatic mutations probably have large effect on grape diversity.

Conclusion

NGS today allows genome wide studies of polymorphism for several individuals. We used a new pipeline of analysis which allows identifying and quantifying SNPs, indels and structural variation between several individuals based on mapping of the sequences on the reference genome but no *do-novo* mapping. We provided a broad description of accumulated somatic mutations during plant life and evaluated their significance in clonal and cultivar diversity generation. Mutations accumulated along life of perennial plants result in clonal diversity and so amplify diversity in individuals generated through sexual reproduction. We identified two types of molecular mechanisms which are responsible for genetic polymorphism: SNP / indel stable mutations which generated a lower polymorphism than structural which for a large part may be eliminated. Few mutations were also observed in exons, which may have a direct impact on phenotype variation.

Acknowledgements

We are grateful to Dr. Anne-Francoise Adam Blondon, Dr. Frédérique Pelsy, Dr. Franc-Christophe Baurens and Dr. Francois Sabot for discussions around this study.

This work was funded by the French Ministry of Research and Higher education and the French Ministry of Food, Agriculture and Fisheries; including a PhD grant from the IFV for GC.

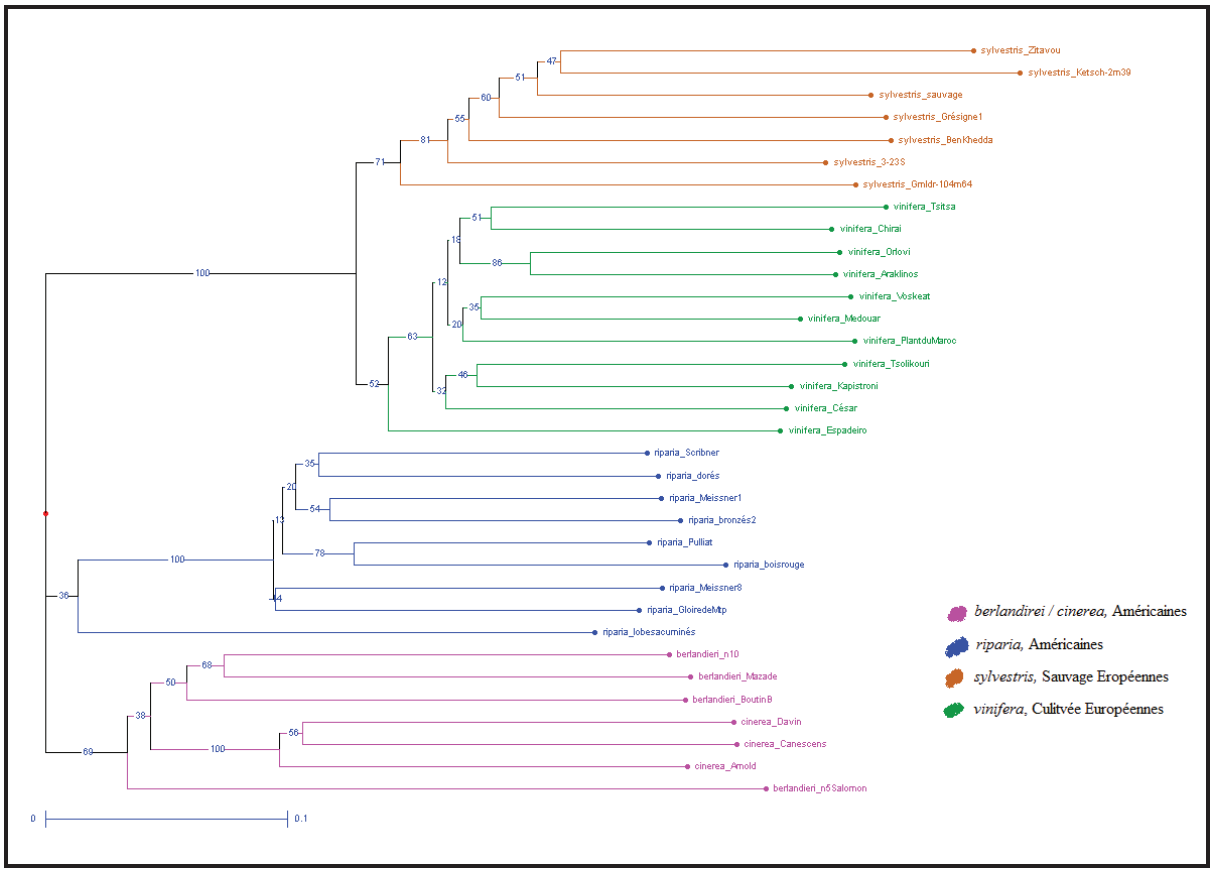


Figure 47 : Arbre phylogénétique de différentes espèces du genre *Vitis* à partir de quatre éléments transposables.

5- Etudes complémentaires : polymorphisme d'insertion généré par les éléments transposables

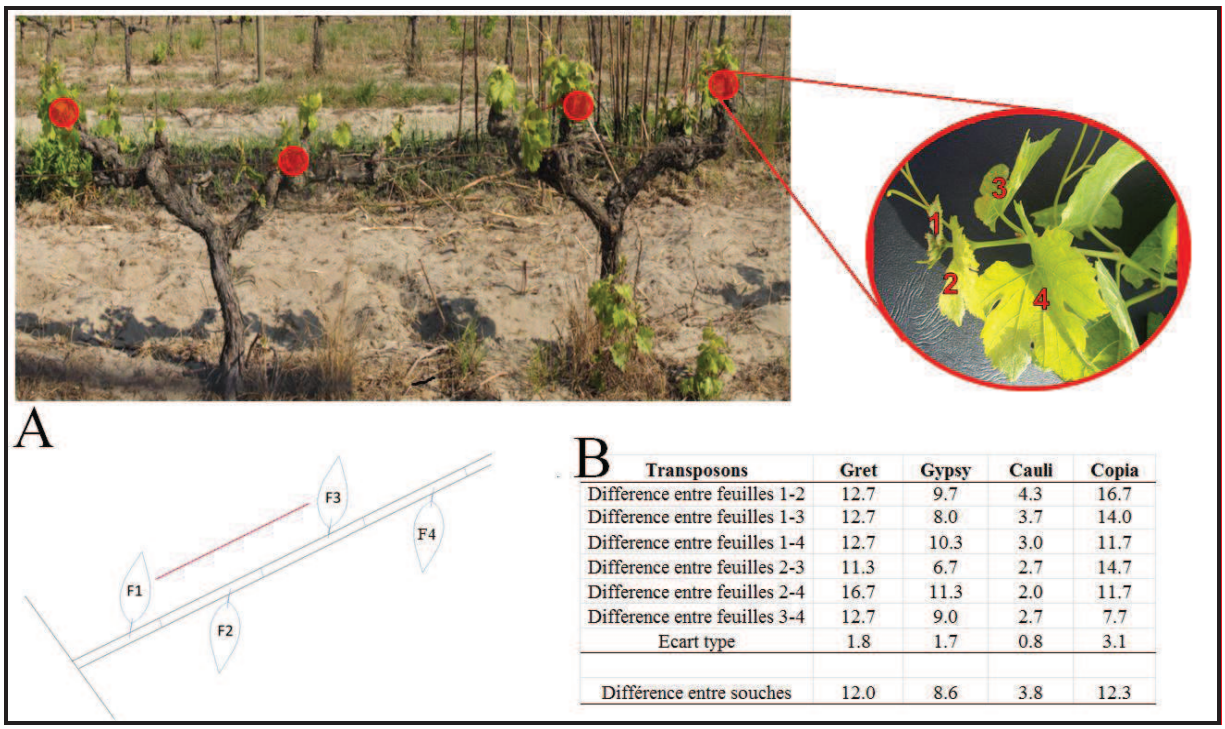
Les deux études précédentes ont montré que l'activité des éléments transposables était très importante dans le contexte de la reproduction végétative chez la vigne. Nous avons effectué deux études complémentaires au cours desquelles nous avons observé l'activité d'insertion de quatre éléments transposables (*Gret-1, Copia-10, Gypsy-19* et *Caul*-1). Ces quatre éléments ont été sectionnés en fonction du haut niveau de polymorphisme d'insertion identifié au sein des clones de Pinot (Cf. Chapitre 3, Section 5) et nous avons observé le polymorphisme d'insertion dans un échantillon diversifié de différentes espèces du genre *Vitis* et au sein d'un même clone.

5.1-Polymorphisme d'insertion de quatre éléments transposables dans le genre Vitis

Au cours de cette thèse j'ai eu l'occasion de co-encadrer un étudiant de Licence 3ème année : Pierre Bourguet. Je présente dans ce paragraphe une partie des résultats qu'il a obtenus au cours de son stage.

Le polymorphisme d'insertion de quatre éléments transposables : *Gret-1, Copia-10, Gypsy-19* et *Caul-1* a été étudié sur un échantillon composé de quatre core-collections chacune représentative de la diversité d'une espèce du genre *Vitis* (*berlandieri / cinerea* ; *riparia* ; *vinifera* subsp.*vinifera*; *vinifera* subsp. *sylvestris ;* Cf. Annexe-6). Ces quatre core-collections ont été définies à partir de données de génotypage de 20 SSRs répartis sur le génome de la vigne. La méthodologie S-SAP a été utilisée pour l'étude des sites d'insertion de ces quatre éléments transposables. A partir des profils obtenus, un arbre phylogénétique a été réalisé (Figure 47 ; méthode neighbor-joining avec 1000 bootstraps) à l'aide du logiciel Darwin (Perrier & Jacquemond Collet, 2006).

Les résultats concordent avec la structure du genre *Vitis* obtenue à l'aide d'autres types de marqueurs moléculaires SNPs et SSRs Péros *et al*, (2010), (Bacilieri, Com. Pers.). Cette étude



B

Transposons	Gret	Gypsy	Cauli	Copia
Difference entre feuilles 1-2	12.7	9.7	4.3	16.7
Difference entre feuilles 1-3	12.7	8.0	3.7	14.0
Difference entre feuilles 1-4	12.7	10.3	3.0	11.7
Difference entre feuilles 2-3	11.3	6.7	2.7	14.7
Difference entre feuilles 2-4	16.7	11.3	2.0	11.7
Difference entre feuilles 3-4	12.7	9.0	2.7	7.7
Ecart type	1.8	1.7	0.8	3.1
Différence entre souches	12.0	8.6	3.8	12.3

Figure 48 : Polymorphisme d'insertion au sein d'un clone (3 souches). A) Exemple d'échantillon prélevé sur une souche, la feuille 1 correspond à la feuille mise en place le plus tardivement, la feuille 4 correspond au contraire à la feuille la plus récemment mise en place. Table B) Moyennes des polymorphismes d'insertion observés entre trois souches du clone de Pinot noir n°777 pour différents organes prélevés sur un même rameau.

préliminaire montre que ces quatre éléments transposables engendrent du polymorphisme d'insertion au moins depuis la divergence des différentes espèces du genre *Vitis* (2,5 x 10^6 années ; Péros *et al*., 2010). Une partie au moins du polymorphisme d'insertion généré par ces quatre éléments s'est donc accumulée au cours temps.

5.2- Polymorphisme d'insertion de quatre éléments transposables au sein d'un même individu

Nous avons mesuré le polymorphisme d'insertion de ces mêmes quatre éléments transposables (*Gret-1, Copia-10, Gypsy-19 et Caul-1),* au sein d'un même individu afin d'estimer leur niveau de transposition.

Trois souches du clone Pinot noir ENTAV-INRA® n°777 cultivées au domaine de l'Espiguette ont été sélectionnées. Les feuilles une à quatre de deux rameaux par souche ont été récoltées et l'ADN extrait (Figure 48). Une approche S-SAP a permis d'identifier les différentes insertions au sein de ces différents organes. Les analyses comparatives des profils obtenus avec l'ADN de chacune des feuilles d'un même rameau montrent de nouvelles insertions de ces éléments, indiquant une activité d'insertion. Cependant ces insertions ne suivent pas une logique d'accumulation, il n'y a pas plus de polymorphisme entre deux feuilles d'un même rameau qu'entre deux souches (Figure 48). Ces résultats suggèrent que la majorité des polymorphismes d'insertions qui se produisent au sein d'un individu sont éliminés rapidement.

Les feuilles ont été prélevées en même temps ne permettant pas l'analyse d'une dynamique d'insertion. Une prochaine étude de ces quatre éléments transposables utilisés va être effectuée en collaboration avec le Docteur Mirouze M. (IRD UMR DIADE). Dans un premier temps, l'activité de transposition de ces éléments va être mesurée et selon les résultats, le suivi de certaines insertions générées par ces éléments sera effectué au cours du temps.

6-Synthèse du polymorphisme clonal chez plusieurs cépages

Un panorama approfondi des polymorphismes qui se sont accumulés au cours la vie de la plante a été dressé chez onze clones répartis entre quatre cépages différents. Le nombre de polymorphismes moléculaires entre deux clones n'est pas significativement différent entre les quatre cépages. Ces résultats suggèrent qu'il n'y pas d'effet cépage sur les mécanismes régulant la fréquence d'apparition des polymorphismes chez les clones étudiés dans ces conditions culturales.

Le nombre de polymorphismes de types SNPs et indels accumulés entre deux clones est faible par comparaison au nombre de polymorphismes de types variations structurales. Cependant la différence du nombre de polymorphismes structuraux entre deux clones et entre deux cépages est faible, suggérant qu'une partie de ces variations ne sont pas transmises aux descendants, qu'ils soient issus de la reproduction sexuée ou de la multiplication végétative. Une partie de ces mutations peuvent cependant être à l'origine de la diversité phénotypique observée. Des travaux additionnels doivent permettre de valider cette implication.

Nous avons pu associer une partie des variations structurales avec l'activité des éléments transposables connus (34%) chez la vigne. Cependant, l'annotation des éléments transposables dans le génome de la vigne est encore partielle et il fort probable que l'implication des éléments transposables dans les variations structurales soit largement sous-estimée. Des travaux visant à identifier l'ensemble des différents éléments transposables compris dans le génome de la vigne sont en cours (N. Choisne, Com. Pers.). L'étude effectuée à partir de quatre éléments transposables chez le genre *Vitis* et au sein d'un même individu, nous a permis d'étudier leur activité d'insertion et leur impact sur la structuration de la diversité. Les résultats obtenus indiquent que ces quatre éléments sont encore très actifs. Ils permettent cependant de structurer la diversité du genre *Vitis* comme d'autres types de marqueurs (SNP, SSR) avec des fréquences de mutations inférieures.

La comparaison du nombre d'insertions polymorphes entre les différentes feuilles d'un même rameau montre que ces insertions ne suivent pas une logique d'accumulation suggérant que la majorité d'entre elles est éliminée rapidement.

Chapitre 5, Discussion générale, perspectives et conclusions

1-Introduction

La diversité clonale a une importance majeure pour les viticulteurs. C'est généralement le seul moyen leur permettant une amélioration variétale sans changer de cépage. Les clones possédant des caractères ampélographiques d'intérêts sont sélectionnés et préservés dans des conservatoires. Ils sont par la suite, étudiés dans différentes conditions et enfin agréés s'ils présentent des valeurs agronomiques et technologiques différentes de ceux déjà certifiés. La sélection clonale a ainsi permis, depuis 45 ans, d'obtenir des gains très significatifs pour la filière viti-vinicole. Par exemple, l'utilisation du clone Pinot noir ENTA-INRA n°777 permet une production de grande qualité.

L'objectif de cette thèse était de comprendre l'origine de la diversité phénotypique clonale. L'hypothèse la plus parcimonieuse permettant de l'expliquer est l'accumulation au cours du temps de mutations. Nous avons dressé un panorama le plus exhaustif possible des différents polymorphismes moléculaires. Pour comparer la diversité moléculaire entre les clones nous avons séquencé le génome de clones de Pinot, de Syrah, de Grenache et de Sultanine à l'aide de séquenceurs de nouvelle génération. Nous avons étudié exclusivement les mutations modifant la séquence du génome. D'autres études proposent une origine de type épigénétique de la diversité clonale (Schellenbaum *et al*., 2008). Les variations de juvénilité pourraient par exemple être expliquées par des variations épigénétiques. Cependant les approches NGS pour des études dites « épigénomiques » étaient encore en phase d'optimisation lors de la production de données, elles nécessitent la mise en place de protocoles particuliers de préparation de librairies (Linnarsson, 2010). Les plateformes de génotypage avec lesquelles nous avons collaboré ne fournissent d'ailleurs pas encore ce service.

La comparaison des génomes de différents clones nous a permis de dresser l'inventaire des polymorphismes génétiques accumulés lors de la vie du cépage. Nos résultats montrent que les variations structurales, dont au moins 34% sont générées par l'activité des éléments transposables connus sont la source majeure du polymorphisme clonal. Elles ont donc été étudiées plus en détails durant ces trois années de recherche afin d'évaluer la possibilité de les utiliser dans une méthodologie permettant l'identification des clones et l'analyse de leur diversité.

2-Comparaison des technologies 454 Titanium et Illumina HiSeq 2000

Nous avons utilisé les séquenceurs de nouvelle génération pour séquencer le génome de plusieurs clones de différents cépages. Ces nouvelles technologies sont très récentes et évoluent très vite. Durant cette thèse, nous nous sommes adaptés à ces évolutions. La sélection de nos travaux comme projets pilotes sur la plateforme Génotoul de l'INRA Toulouse a facilité l'opportunité d'être au plus près des dernières technologies. Nous avons ainsi eu accès aux technologies 454 de Roche Version Titanium (2009) et HiSeq 2000 d'Illumina v2 (2011). Nous pouvons donc comparer ces deux technologies à plusieurs niveaux : i) la production des données, ii) la qualité des génomes reconstruits par alignement, iii) les polymorphismes identifiés.

2.1-Production des données

Ces deux technologies diffèrent sur la quantité et la longueur des séquences produites. De façon générale, le 454 de Roche produit un petit nombre de séquences mais celles-ci sont de grandes tailles. Au contraire, l'Illumina produit des séquences de tailles plus réduites mais en plus grand nombre. En 2009, nous avions accès au 454 Titanium qui produisait des séquences de 400 bases ou à l'Illumina GAII qui produisait des séquences de 50 bases. Ainsi, même si le GAII produisait environ quatre fois plus de données (1,5 Gb pour le GAII contre 400 Mb pour le 454 Titatium) nous avons préféré utiliser le 454. En effet les séquences d'une longueur d'au moins 100 bases permettent d'effectuer plus facilement la reconstruction du génome par alignement (Flicek & Birney, 2009), en particulier pour des génomes au fort polymorphisme comme la vigne (Velasco *et al.*, 2007). De plus notre projet s'inscrivait dans une recherche et une quantification sans *a priori* des types de polymorphismes moléculaires différenciant le génome de deux clones. Seules des séquences de grandes tailles permettaient d'avoir accès aux variations structurales. En 2011, la technologie HiSeq 2000 v2 d'Illumina pouvait fournir en moyenne 20 Gb de données par ligne avec des séquences de 100 bases en "Paired-End". Nous avons donc alors choisi, en 2011, d'utiliser l'Illumina HiSeq 2000 pour effectuer nos séquençages. Ceci nous a permis d'avoir un grand nombre de données avec de séquences de plus petites tailles mais en "Paired-End" donnant également accès aux variations structurales.

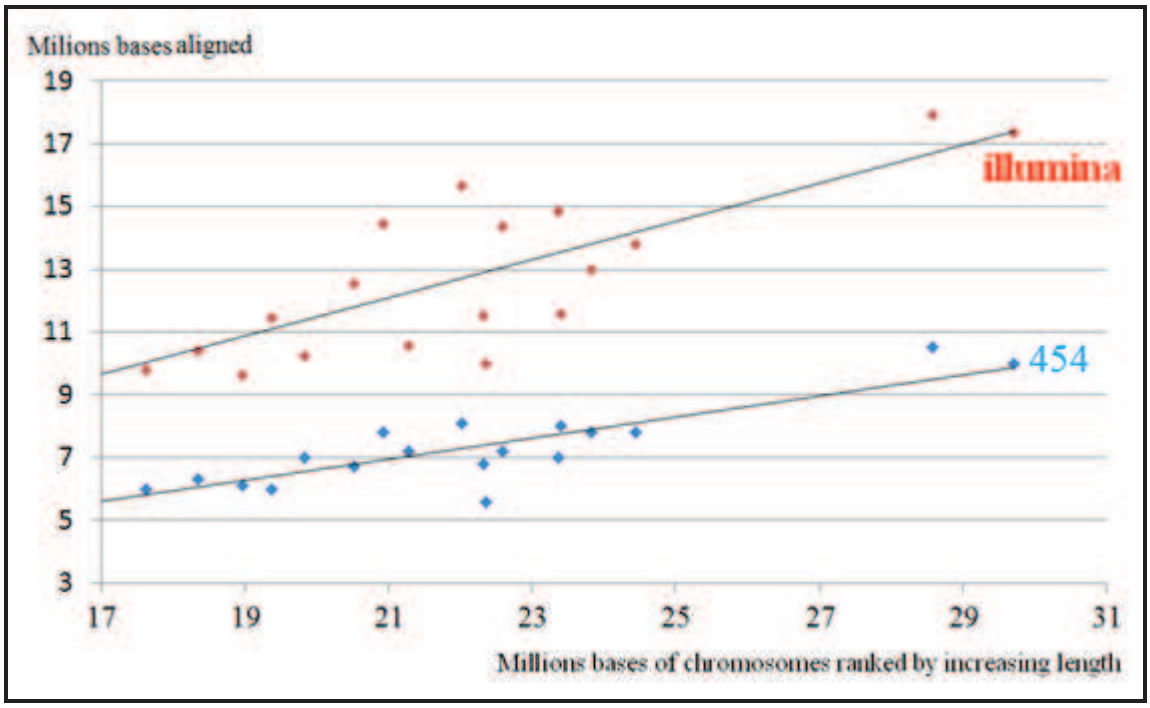


Figure 49 : Distribution des séquences alignées obtenues en Illumina et en 454 pour le clone de Pinot noir n°777 en fonction de la taille des chromosomes rangés par ordre croissant.

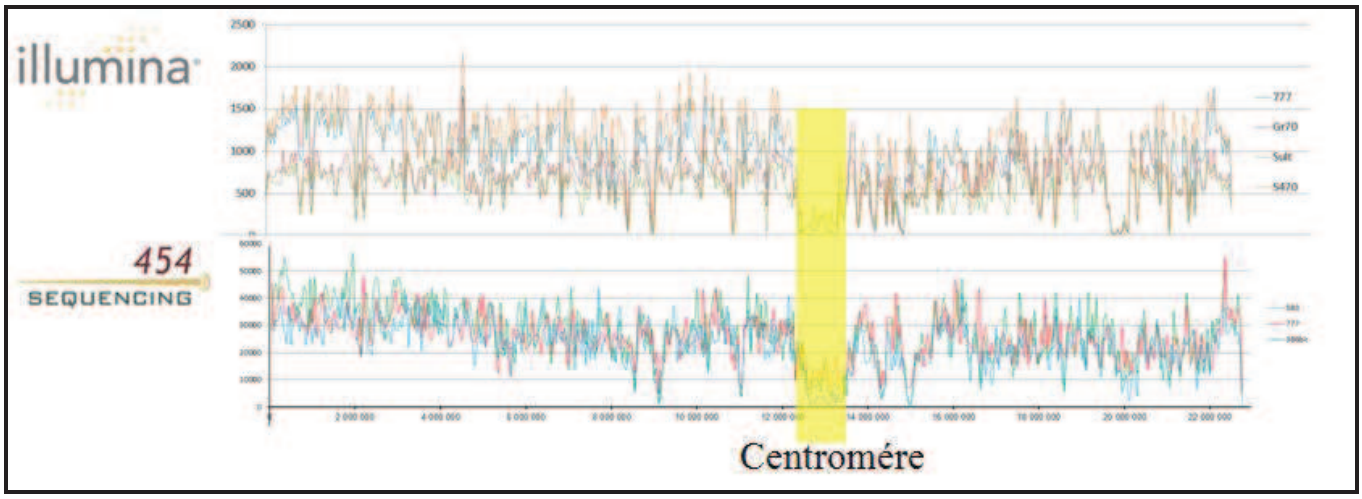


Figure 50 : Distribution des séquences alignées obtenu en Illumina et en 454 pour le clone de Pinot noir n°777 sur le chromosome 1.

2.2-Qualité de la reconstruction du génome

Les séquences obtenues par les NGS ont été utilisées pour reconstruire le génome de l'individu. Elles sont dans un premier temps filtrées afin de ne conserver que les séquences de bonne qualité qui seront utilisées pour l'alignement sur le génome de référence. Dans cette étude, nous avons considéré qu'une séquence était correctement alignée si elle ne pouvait être positionnée qu'à un locus unique sur le génome de référence. Pour les deux technologies utilisées, 57% des séquences en moyenne (62% (σ=9) pour le 454 et 53% (σ=12) pour l'Illumina) ont été alignées sur la référence. Dans les deux cas, la couverture nucléotidique de chaque chromosome a toujours été proportionnelle à leur taille (Figure 49). Ces deux technologies permettent de séquencer l'ensemble du génome sans *a priori*. La répartition des séquences sur un chromosome est également identique pour ces deux technologies, les zones riches en éléments répétés sont peu couvertes, et en particulier la zone centromérique (Figure 50). La composition en GC, CnG et CpG est similaire entre les génomes obtenus en 454 ou en Illumina. Les génomes reconstruits par les technologies 454 ou Illumina sont donc de qualité identique. Cependant, nous n'avons pas la même couverture, avec l'Illumina nous couvrons environ 50% du génome de la vigne avec une profondeur de 10X ou plus, alors que nous ne couvrons que 1% du génome de la vigne avec une profondeur de 6X ou plus pour le 454.

Avec les méthodes d'alignement utilisées, nous avons difficilement accès à la partie répétée du génome de la vigne (41%, d'après Jaillon *et al*. (2007)). Cette partie du génome ne peut pas être reconstruite par alignement ce qui explique probablement qu'une grande partie des séquences issues de nos re-séquençages soient non alignées. Une partie du polymorphisme généré par les éléments répétés a cependant pu être étudiée. Par contre, cette étude ne concerne que les sites d'insertions et non le polymorphisme interne aux éléments transposables. Pour étudier les régions répétées, il faut privilégier les méthodes d'assemblage plutôt que les méthodes d'alignement. Les méthodes d'assemblage *de novo* à partir des séquences NGS nécessitent de grandes capacités informatiques. Actuellement, l'analyse s'oriente vers un compromis de ces approches appelé « assemblage *de novo* guidé » qui consiste à ne reconstruire que les zones présentant une importante insertion ou délétion par rapport à la référence. Pour m'aider dans la mise en place du « Bacchus pipeline », j'ai eu l'occasion d'encadrer Vincent Maillol, alors stagiaire en licence. Il est désormais recruté au sein de l'équipe et travaille actuellement dans l'amélioration du Bacchus pipeline.

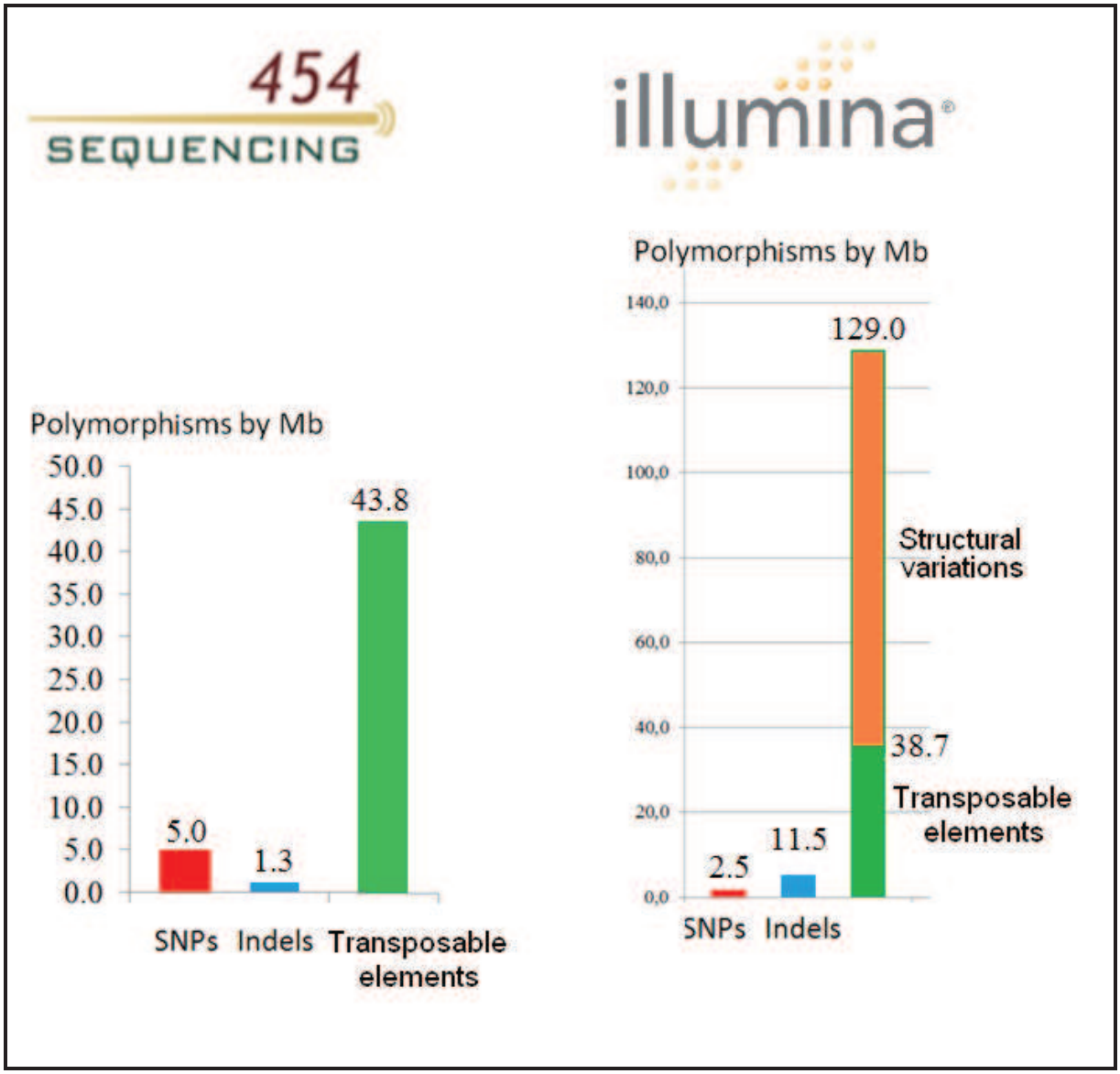


Figure 51 : Comparaison des moyennes des différents types de polymorphismes identifiés entre deux clones avec les technologies 454 et Illumina.

2.3-Polymorphismes identifiés

Les génomes reconstruits des différents clones ont été comparés afin d'identifier les polymorphismes moléculaires. Les polymorphismes de types SNPs et indels (de moins de 20 bases, Cf. Chapitre 4, Section 4) ont pu être détectés à la fois dans les génomes reconstruits en 454 et en Illumina. Les indels compris entre 20 et 100 bases correspondant à des séquences répétées (en particulier les motifs SSR de 100 bases) ne peuvent être détectés qu'avec des séquences de grandes tailles et leur identification est donc limitée aux génomes reconstruits en 454. Les variations structurales générées par les éléments transposables connus chez la vigne ont pu être observées en 454 et en Illumina. De plus, la technologie "Paired-End" utilisée en Illumina nous a permis d'observer l'ensemble des variations structurales et pas seulement celles générées par l'activité des éléments mobiles connus chez la vigne.

Ces deux technologies apparaissent comme complémentaires. La technologie 454 permet cependant, avec ses séquences plus longues, d'identifier plus facilement du polymorphisme. Il est à noter que la technologie "Paired-End" en 454 est maintenant disponible. L'avantage d'avoir des séquences de grandes tailles en "Paired-End" est de faciliter l'identification des différents types de variations structurales. En effet, avec des séquences de 100 bases, il est difficile d'identifier avec précision l'élément transposable inséré sans effectuer des étapes d'assemblage supplémentaires. L'avantage incontestable du HiSeq 2000 d'Illumina est la grande quantité de données produites lors des "runs" de séquences (20 Gb par ligne) alors que le 454 produit 350 Mb, avec une couverture qui s'est avérée limitée (1% du génome).

Une partie des polymorphismes a été validée, selon différents paramètres de validations, pour avoir une estimation du taux de faux positifs. Dans un premier temps, nous avons sélectionné les polymorphismes qui ont été séquencés au moins 3 fois en 454 (3X). 16 SNPs sélectionnés au hasard ont été séquencés en Sanger et seul un SNP s'est avéré réel. Ces erreurs étaient principalement dues au fait qu'un seul des deux haplotypes de l'individu avait été séquencé. En ne sélectionnant que les SNPs qui ont une profondeur de 6X, 19 SNPs ont été identifiés et, dans ce cas, seuls deux SNPs avèrent être des faux positifs. Plus la profondeur est importante, plus le taux d'erreur est faible, et avec une profondeur supérieure à 10X le taux d'erreur peut être considéré comme nul (Harismendy *et al.*, 2009). Les résultats obtenus avec le 454 montrent un nombre de polymorphismes par Mb légèrement supérieur à celui obtenu en Illumina pour les polymorphismes de type SNPs et éléments transposables (Figure 51). Le nombre de faux positifs plus élevé provenant des données obtenues en 454

explique cette différence. Au contraire, le nombre de polymorphismes de type indels obtenu avec le 454 est plus faible que celui obtenu en Illumina. Afin de limiter les faux positifs, de nombreux polymorphismes indels contenus dans les homopolyméres n'ont pas été considérés en 454 ce qui pourrait expliquer cette différence.

2.4-Les NGS révolutionnent le génotypage

Avec les nouvelles générations de séquenceurs, un génome entier peut être maintenant séquencé en quelques jours, alors qu'auparavant cela demandait plusieurs années. Les technologies évoluent très vite et le volume de données augmente exponentiellement. Ma thèse est à l'image de cette évolution rapide. En moins de deux ans, le volume de données que nous avons pu obtenir a été multiplié par 50. La technique Illumina nous a permis d'être beaucoup plus exhaustifs et la technologie "Paired-End" a permis d'identifier une partie des variations structurales. Cependant, la taille des séquences reste faible comparée au 454, ce qui limite l'identification de certains polymorphismes. Le traitement de ces montagnes de données nécessite par ailleurs la mise en place d'infrastructures informatiques et de nouvelles méthodes d'analyses. L'automatisation de l'analyse des données est devenue une obligation. Ainsi, les logiciels permettant d'analyser ce type de données évoluent également très vite et il est souvent difficile de choisir entre deux logiciels. Dans ce travail de thèse nous avons mis en place un pipeline constitué de logiciels disponibles et de scripts développés en interne. Ce pipeline est disponible sous l'interface Galaxy et peut être utilisé sur n'importe quel type de séquences NGS. Il permet de reconstruire un génome par une méthode d'alignement et de le comparer à d'autres génomes afin d'identifier tous les types de polymorphismes. Ce pipeline est en constante évolution et de nouveaux modules seront ajoutés pour permettre une analyse plus exhaustive, en particulier pour l'assemblage *de-novo*.

3-Importance des mutations à l'origine de la diversité de la vigne

Un clone est généralement défini comme la copie identique de l'individu dont il est issu. Cependant au cours de la vie des plantes, les mutations s'accumulent. Ainsi plus la plante est âgée et a subi de cycles de multiplication végétative, plus elle accumule de mutations et donnera naissance à des individus diversifiés (Klekowski, 1998). Dans cette thèse

nous avons quantifié et décrit les différentes mutations accumulées au cours de la vie du cépage et transmises par multiplication végétative.

3.1-Chimérisme et impact des mutations

Un clone est issu de la régénération d'un tissu sélectionné chez une plante matrice. La propagation des mutations dans la plante mère dépend du lieu où celles-ci apparaissent. Si elle apparait dans une cellule initiale d'un bourgeon apical, elle sera transmise à tout le rameau issu de ce bourgeon et a donc plus de chances d'être transmise à la descendance végétative que si elle n'apparait par exemple que dans une seule feuille. Seules les mutations accumulées dans la partie de la plante bouturée seront transmises au nouveau clone. La vigne, comme la plupart des plantes, est composée d'au moins deux lignées cellulaires (L1 et L2). La lignée cellulaire L1 donne naissance aux tissus épidermiques alors que la L2 donne naissance aux tissus vasculaires, parenchymateux et aux gamètes. De par la structuration des méristèmes, une mutation peut apparaitre soit dans la L1, soit dans la L2. Elle peut donc donner naissance à des chimères. Les données obtenues sur les 11 clones montrent que tous les individus qui ont été analysés sont chimériques du fait de la présence de positions SNPs et/ou indels avec trois allèles. Lors de la multiplication végétative traditionnelle, les deux couches cellulaires sont à la base d'un nouveau clone. La reproduction sexuée permet, quant à elle, la transmission d'une partie des allèles de l'individu. Les gamètes étant issus de la lignée cellulaire L2, seules les mutations accumulées dans cette couche peuvent être transmises à la descendance. L'accumulation de mutations au cours de la vie de la plante participe donc en partie aussi à la diversité des cépages.

3.2-Fréquence d'accumulation des mutations

De façon générale, la fréquence des mutations est plus forte chez les plantes annuelles que chez les plantes pérennes (Klekowski & Godfrey, 1989). De récentes études moléculaires (Petit & Hampe, 2006) tendent à attester que les arbres accumulent moins de mutations par unité de temps que ne le font les plantes annuelles.

Une partie de ces polymorphismes est générée par l'activité des éléments transposables connus chez la vigne (34%). Cette proportion pourrait augmenter lorsque nous aurons accès à une meilleure connaissance de l'ensemble des éléments transposables présents dans le génome de la vigne. Les insertions générées par l'activité des éléments transposables ont une

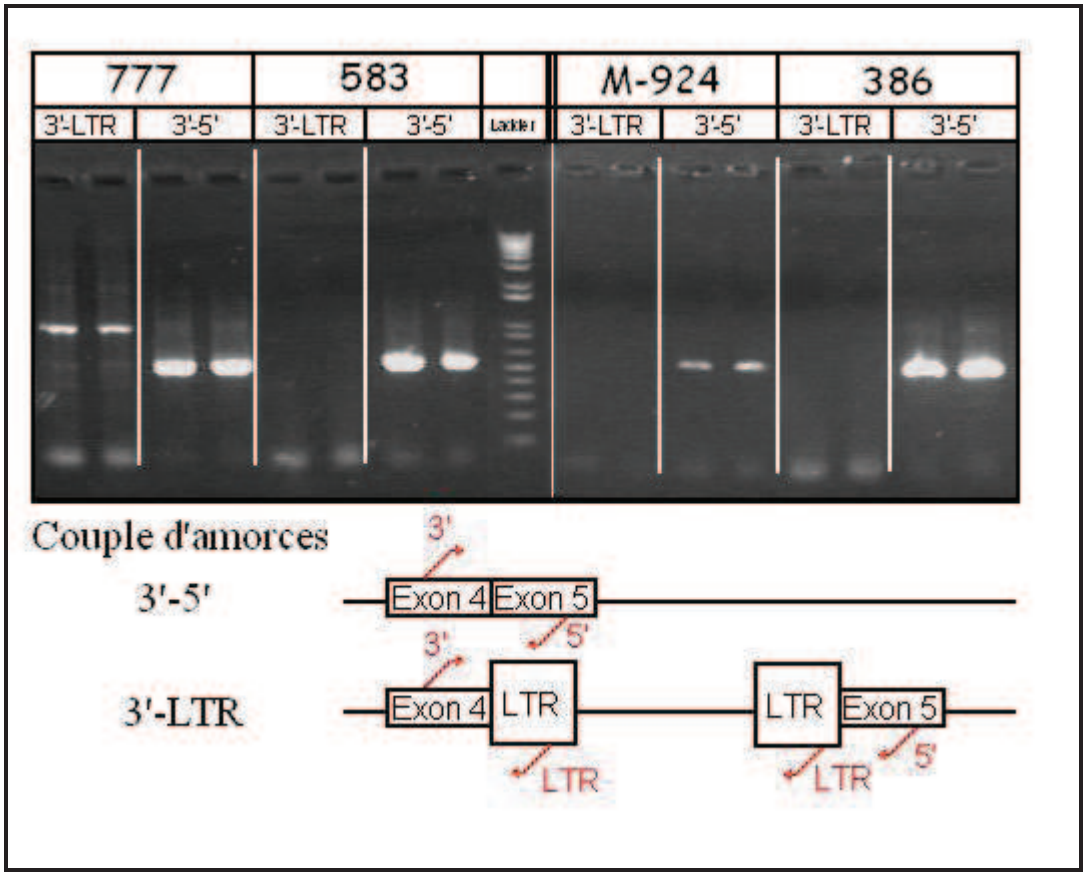


Figure 52. Validation du polymorphisme d'insertion généré par *Copia-34* dans le gène GSVIVT01013391001 par une approche PCR. L'expérience a été doublée.

durée de vie très variable dans le génome. La comparaison du nombre d'insertions polymorphes entre différents organes d'un même individu montre que ces insertions sont très fréquentes, mais ne suivent pas une logique d'accumulation suggérant que la majorité d'entre elles est éliminée rapidement (Cf. Chapitre 4). Les mécanismes d'élimination des rétrotransposons les plus couramment observés engendrent des solo LTR (Devos *et al.*, 2002). Avec l'approche S-SAP utilisée, la distinction entre la présence d'un retrotransposon entier ou d'un solo LTR à un locus est impossible. Les resultats obtenus suggèrent une élimination totale ou un fort réarrangement de l'insertion des rétrotransposons ce qui est en accord avec des études théoriques (Gray, 2000).

Les polymorphismes de type SNP / indel ont une fréquence d'apparition beaucoup plus faible que celle des variations structurales (estimé à 10^{-8} par génération pour les SNPs (Nachman & Crowell, 2000) et à 10^{-4} par génération pour les variations structurales (Inoue & Lupski, 2002)). Cependant la fréquence de réversion des SNPs par rapport aux variations structurales est aussi nettement plus faible. Ces mutations apparaissent généralement lors d'une erreur de réplication ou de réparation de l'ADN et ne sont pas éliminées. Elles s'accumulent donc dans le génome et peuvent être transmises à la descendance selon leur localisation.

3.3-Impact des mutations sur le phénotype et la sélection assistée par marqueurs

Certaines mutations sont responsables de variations phénotypiques. L'identification de ces mutations et leur association avec un phénotype permet la mise en place de marqueurs moléculaires pouvant être utilisés dans les programmes de sélection. Les mutations identifiées entre les différents clones au cours de cette thèse permettront l'établissement de polymorphismes candidats jouant un rôle dans les variations phénotypiques.

De façon générale, une mutation a plus de chance d'affecter le phénotype si elle est localisée dans un exon. Quelques exemples de mutations affectant le phénotype situées dans les exons ont été identifiés chez la vigne (Fournier-Level *et al.*, 2009; Emanuelli *et al.*, 2010). Au cours de cette thèse une mutation candidate a été identifiée. Une insertion a en effet été localisée dans le 4éme exon du gène GSVIVT01013391001 exclusivement chez le clone Pinot noir ENTAV-INRA® n°777 parmi les quatre clones de Pinot étudiés. Ce polymorphisme d'insertion a été généré par l'élément *Copia-34* et a été confirmé par une

approche PCR sur les quatre clones (Figure 52), seul l'individu n°777 est hétérozygote et possède l'insertion. L'homologue le plus proche de ce gène code pour une enzyme responsable de la synthèse du phosphatidylinositol, un phospholipide de la membrane plasmique (identité protéique à 33% chez *Arabidopsis*). La validation fonctionnelle de l'impact de ce polymorphisme sur le phénotype du clone n°777 n'a pas été effectué au cours cette thèse.

D'autres mutations situées dans les régions régulatrices peuvent aussi affecter le phénotype. Par exemple, une insertion du retrotransposon *Hatvine1* dans la région régulatrice du gène *VvTFL1A* entraine sa surexpression et provoque une sur-ramification de la grappe (Fernandez *et al.*, 2010) ou encore l'insertion de l'élément *Gret-1* dans le promoteur du gène *VvMybA1* provoque des différences de couleurs de baies (Kobayashi *et al.*, 2004). Ces mutations situées en amont d'un gène sont fréquentes mais seule une minorité d'entre elles affectent son expression. L'impact de ces mutations candidates est difficile à valider (nécessité de faire des croisements ou des approches de transgénèse). La quantification de ces mutations en amont de gènes est en cours.

Au cours de cette thèse, nous avons étudié la diversité des clones de Pinot par une approche S-SAP avec quatre éléments transposables (Cf. Chapitre 2). Nous avons tenté d'associer ces données de génotypage avec les données phénotypiques dont nous disposions. L'instabilité des marqueurs S-SAP et le peu de données phénotypiques disponibles ne nous a permis que d'observer une structuration très limitée de la diversité des clones de Pinot en fonction des caractères quantitatifs comme la couleur et la villosité. Des études complémentaires avec des données phénotypiques quantitatives permettraient certainement de mieux comprendre la structuration de la diversité des clones.

3.4-Conclusion

Ces travaux ont mis en évidence l'accumulation de mutations entre les génomes des clones pouvant affecter le phénotype. Pour la première fois nous avons décrit et quantifié les différents polymorphismes moléculaires existants dans un contexte de reproduction végétative. Deux types de mutations ont pu être identifiés, des mutations avec une fréquence d'apparition faible mais relativement stables (SNPs et indels) et des mutations fréquentes mais fortement révertantes (variations structurales). Seules les mutations modifiant la séquence du génome ont été étudiées ici, mais les mutations épigénétiques peuvent également

être responsables d'une partie de la variation clonale. Ces mutations ont une régulation complexe, sont instables et dépendantes de l'organe et de l'environnement. Certaines études ont montré une corrélation entre l'activité des éléments transposables et la présence de ces mutations épigénétiques, suggérant un lien étroit entre ces deux éléments (Mirouze *et al.*, 2009).

4-Perspectives

4.1-Vers l'identification clonale

L'identification des clones par une méthodologie moléculaire a une grande importance pour les viticulteurs et également pour l'IFV. Elle permettrait d'obtenir l'assurance de l'identité du clone qui peut être, dans certains cas, difficile à différencier avec les seuls caractères ampélographiques. Une méthodologie basée sur les marqueurs microsatellites permet d'identifier les cépages, cependant ces marqueurs ne sont pas suffisamment polymorphes pour permettre d'identifier les clones (Imazio *et al.*, 2002; Riaz *et al.*, 2002; Moncada & Hinrichsen, 2007; Pelsy, 2009; Pelsy *et al.*, 2010).

Nous nous sommes intéressés aux éléments transposables comme marqueurs moléculaires pour différencier les clones car ce sont les éléments qui entrainent le plus de polymorphisme entre les clones. Des résultats précédents avaient montré des possibilités d'identification des clones avec ce type de marqueurs (Wegscheider *et al.*, 2009; Anhalt *et al.*, 2010). Cependant le choix de l'élément transposable est primordial. Des études faites à partir d'un élément transposable *Vine-1* montraient l'absence de polymorphisme d'insertion empêchant la différenciation des clones mais pas des espèces (Labra *et al.*, 2004). Nous avons alors étudié la diversité des clones de Pinot à l'aide de quatre éléments transposables identifiés comme hautement polymorphes. Avec ces marqueurs moléculaires nous avons obtenu un profil unique pour chacun des clones testés. Les données complémentaires obtenues par la comparaison des profils au sein d'un même individu montrent cependant que des éléments trop actifs sont instables et difficilement utilisables pour l'identification des clones. Un compromis entre activité et stabilité devra sans doute être trouvé.

Le polymorphisme SNP et indel entre les clones est faible comparé aux variations structurales, mais ces polymorphismes sont plus stables. Il serait alors peut-être plus judicieux

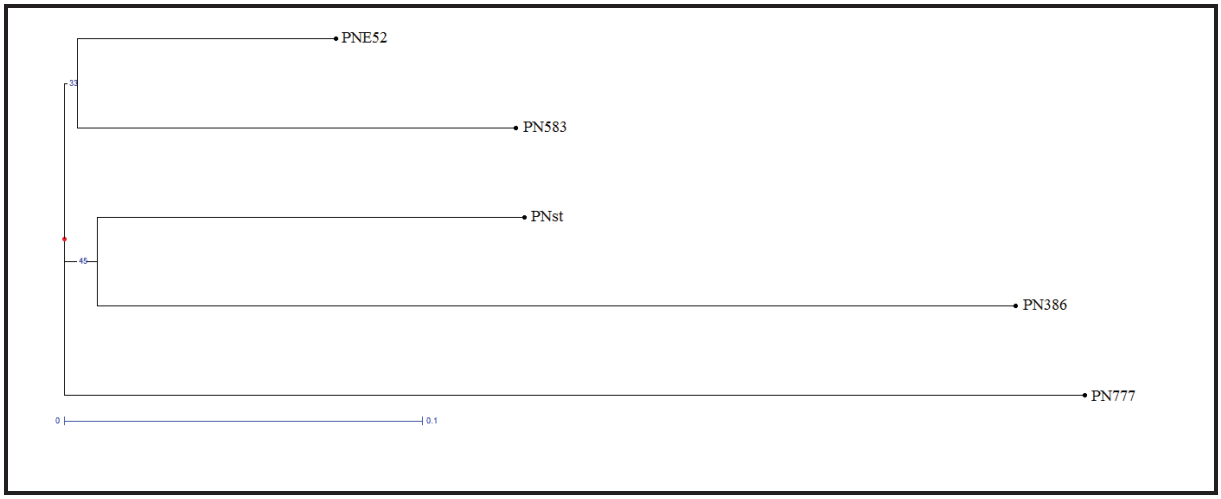


Figure 53 : Arbre phylogénétique obtenu à partir des SNPs polymorphes entre les 5 clones de Pinots séquencés en Illumina HiSeq 2000.

pour l'identification clonale de s'orienter vers ce type de polymorphismes. Nous avons établi un arbre phylogénétique à partir des polymorphismes SNPs détectés entre les cinq clones de Pinot séquencés en HiSeq 2000 (Figure 53, neighbor-joining avec 1000 bootstraps). Cet arbre permet de retracer l'histoire des lignées de ces cinq clones de Pinot. On constate que le Pinot ENTAV-INRA® n°777 descend d'une lignée indépendante des quatre autres clones de Pinot analysés. Cependant, même si cela reste possible, le séquençage complet de tous les clones est à ce jour encore difficile à envisager. Au vu de l'avancée des NGS, il est fort probable que dans quelques décennies un séquençage de génome complet devienne aussi commun que la méthode PCR. La limite sera probablement au niveau de l'analyse bioinformatique. Il nous parait donc plus judicieux à l'heure actuelle d'envisager la création d'une puce SNP permettant de distinguer de nombreux clones en une même analyse.

Etant donné les capacités de séquencage actuelles, tous les clones agréés d'un cépage pourraient être séquencés afin d'établir une liste de SNPs permettant d'identifier tous les clones agréés. Ces SNPs alors identifiés permettront l'établissement d'une puce pouvant être utilisée pour l'identification des clones. De plus, s'il est démontré ultérieurement que certains SNPs sont corrélés à un phénotype, ces puces pourraient aussi être utilisées en sélection assistée par marqueurs.

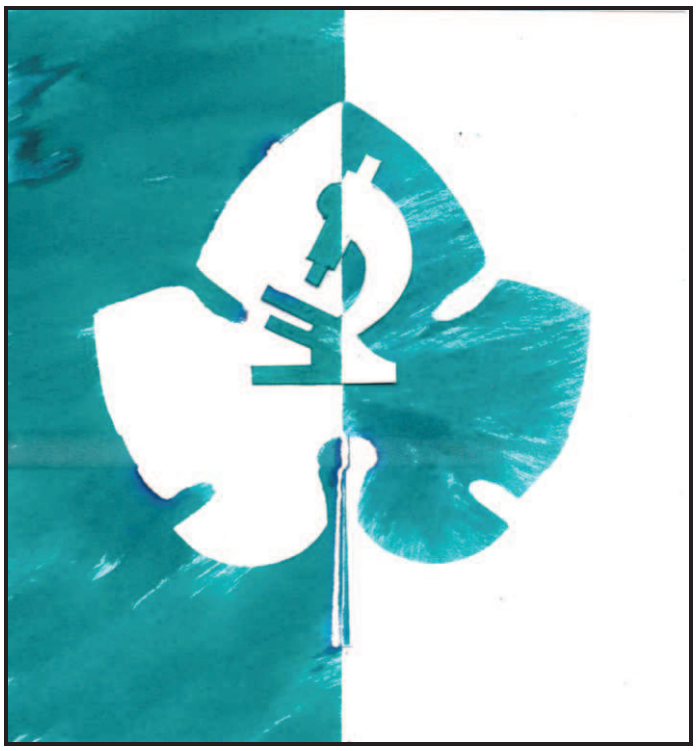
4.2-Vers la sélection assistée par marqueurs

Une des applications plus ou moins directe de ces travaux sera également la mise en place d'une aide à la sélection des clones par marqueurs moléculaires. Les polymorphismes identifiés entre les clones pourront être utilisés afin de définir des marqueurs liés à la variabilité phénotypique ou plus globalement, plusieurs marqueurs bien répartis sur le génome pourront être corrélés à un profil phénotypique clonal (sélection génomique).

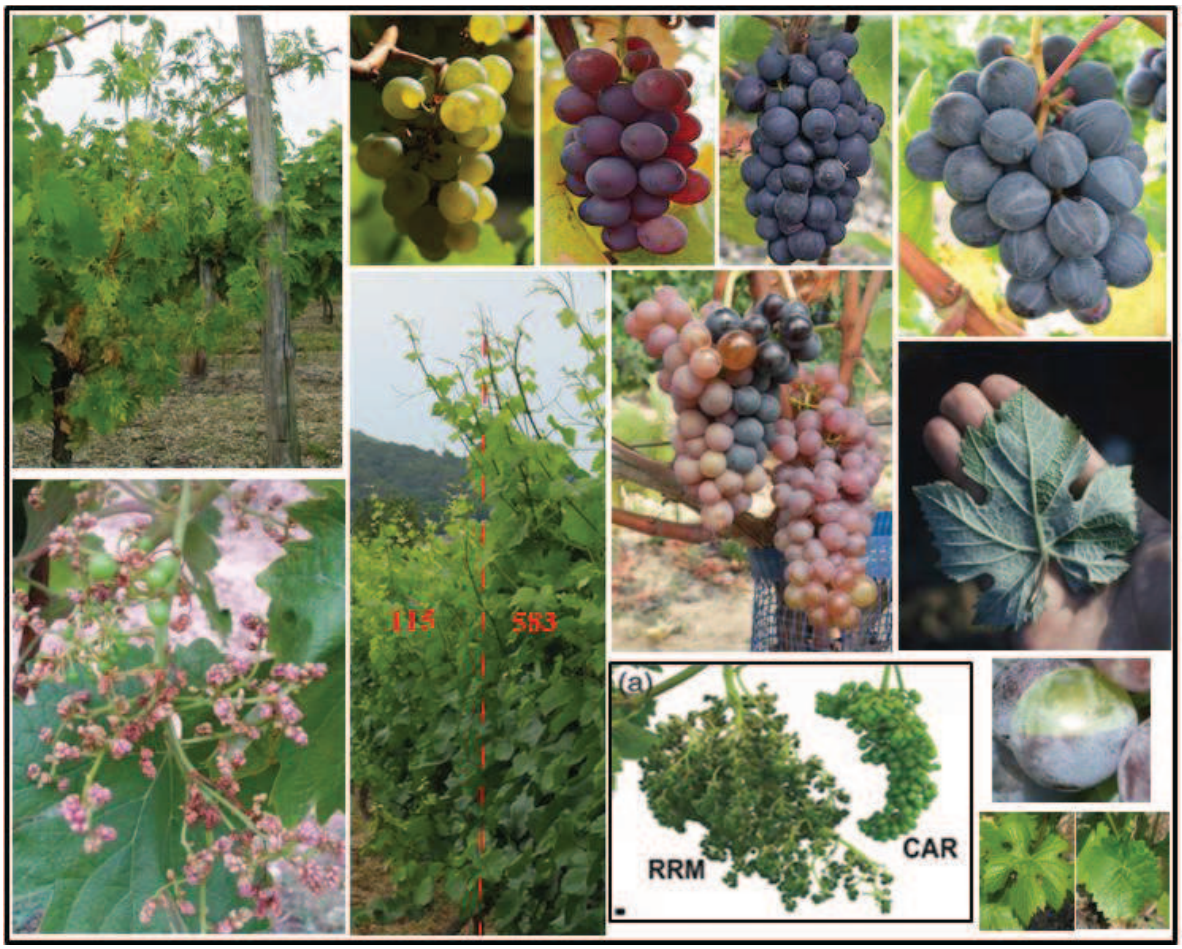
Deux objectifs pourront ainsi être poursuivis: i) mieux caractériser la diversité génétique disponible dans les conservatoires pour mieux cibler la diversité utile qui permettra une orientation différente de la sélection, à l'image de ce qui a été fait pour les conservatoires de clones de Pinot, ii) accélérer les schémas de sélection en utilisant des marqueurs moléculaires permettant de prédire les caractères phénotypiques recherchés (arômes particuliers, taux de sucre plus faible, port érigé,…).

Ces objectifs sont à considérer à moyen terme. Cette thèse, en démontrant en particulier la faisabilité de l'approche et par les polymorphismes qu'elle a déjà permis d'identifier, constitue un premier pas vers cet objectif ambitieux.

L'UMT Génovigne® a été créée dans le but de permettre un transfert des résultats de la recherche vers le monde appliqué. En réalisant cette thèse sur un thème de recherche en amont mais qui peut avoir des retombées concrètes sur un problème important pour la viticulture, je pense avoir contribué à ma façon à ce transfert.



Jean jacques Pajeile


115
583
(a)
RRM
CAR

Références

Adam-Blondon AF, Roux C, Claux D, Butterlin G, Merdinoglu D, This P. 2004. Mapping 245 ssr markers on the vitis vinifera genome: A tool for grape genetics. *TAG Theoretical and Applied Genetics* **109**(5): 1017-1027.

Agius F, Kapoor A, Zhu JK. 2006. Role of the arabidopsis DNA glycosylase/lyase ros1 in active DNA demethylation. *Proceedings of the National Academy of Sciences* **103**(31): 11796-11801.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W. 1997. National center of biotechnology information gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389 - 3402.

Anhalt CM, Crespo Martínez S, Rühl E, Forneck A. 2010. Dynamic grapevine clones an aflp-marker study of the vitis vinifera cultivar riesling comprising 86. *Tree Genetics & Genomes* **Springer-Verlag-online**.

Ansorge WJ. 2009. Next-generation DNA sequencing techniques. *New Biotechnology* **25**(4): 195-203.

Antoni R. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5**(2): 94-100.

Aradhya MK, Dangl GS, Prins BH, Boursiquot JM, Walker MA, Meredith CP. 2003. Genetic structure and differentiation in cultivated grape, vitis vinifera l. *Genet Res* **81**: 179-191.

Atanur SS, Birol Än, Guryev V, Hirst M, Hummel O, Morrissey C, Behmoaras J, Fernandez-Suarez XM, Johnson MD, McLaren WM, Patone G, Petretto E, Plessy C, Rockland KS, Rockland C, Saar K, Zhao Y, Carninci P, Flicek P, Kurtz T, Cuppen E, Pravenec M, Hubner N, Jones SJM, Birney E, Aitman TJ. 2010. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. *Genome Research* **20**(6): 791-803.

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: Causes and consequences. *Nat Rev Genet* **8**(8): 619-631.

Bannert N, Kurth R. 2006. The evolutionary dynamics of human endogenous retroviral families. *Annual Review of Genomics and Human Genetics* **7**(1): 149-173.

Benaglio P, Rivolta C. 2010. Ultra high throughput sequencing in human DNA variation detection: A comparative study on the ndufa3-prpf31 region. *PLos One* **5**(9): e13071.

Benjak A, Boue S, Forneck A, Casacuberta JM. 2009. Recent amplification and impact of mites on the genome of grapevine (vitis vinifera l.). *Genome Biol Evol* **2009**(0): 75-84.

Benjak A, Forneck A, Casacuberta JM. 2008. Genome-wide analysis of the “cut-and-paste” transposons of grapevine. *PLos One* **3**(9): e3107.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**: 6-21.

Blankenberg D, Kuster G, Von,, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2001. Galaxy: A web-based genome analysis tool for experimentalists. *Current protocols in molecular biology*: John Wiley & Sons, Inc.

Boffey SA, Leech RM. 1982. Chloroplast DNA levels and the control of chloroplast division in light-grown wheat leaves. *Plant Physiol.* **69**(6): 1387-1391.

Boss PK, Thomas MR. 2002. Association of dwarfism and floral induction with a grape `green revolution' mutation. *Nature* **416**(6883): 847-850.

Bouquet A. 1982. Origine et évolution de l'encépagement francais à travers les siècles. *Progrès agricole et viticole* **5**(110-121).

Bouquet A, Boursiquot JM. 1999. La sauvegarde des vieux cépages et la création de variétés nouvelles: Une démarche conjointe pour concilier tradition et innovation en france. *PAV* **72**(825-26): 753-761.

Bouquet A, Davis HP, Danglot Y, Rennes C. 1989. Culture in vitro d'ovules et d'embryons de vigne (vitis vinifera l.) appliquée à la sélection de variétées de raisins de table sans pépins. *Agronomie* **9**(6): 565-574.

Boursiquot JM, Audeguin L, Charmont S, Desperrier JM, Dufour MC, Jacquet O, Lacombe T, Leguay M, Moulliet C, Ollat N, Schneider C, Serreno C. 2007. Catalogue des varietes et clones de vignes cultives en france. *Institut Francais de la Vigne et du Vin* **2**.

Boursiquot JM, Dessup M, Rennes C. 1995. Distribution of the main phenological, agronomical and technological characters of vitis-vinifera l. *Vitis* **34**(1): 31 - 35.

Boursiquot JM, Lacombe T, Laucou V, Julliard S, Perrin FX, Lanier N, Legrand D, Meredith C, This P. 2009. Parentage of merlot and related winegrape cultivars of southwestern france: Discovery of the missing link. *Australian Journal of Grape and Wine Research* **15**(2): 144-155.

Boursiquot JM, This P. 1996. Les nouvelles techniques utilisées en ampélographie: Informatique et marquage. *J Int Sci Vigne Vin* **Special issue**: 12-23.

Boursiquot JM, This P. 2000. Essai de définition du cépage. : *PAV* (94):5-7.

Bowen N, Jordan I. 2002. Transposable elements and the evolution of eukaryotic complexity. *Mol Biol* **4**(3): 65-76.

Bowers JE, Boursiquot JM, This P, Chu K, Johansson H, Meredith CP. 1999. Historical genetics: The parentage of chardonnay, gamay, and other wine grapes of northeastern france. *Science* **285**: 1562-1565.

Bräutigam A, Gowik U. 2011. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology* **12**(6): 831-841.

Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, Audeguin L, Boursiquot JM, This P. submitted. Transposable elements are a major cause of somatic polymorphism in vitis vinifera l. . *PLoSONE*.

Carrier G, Santoni S, Rodier-Goud M, Canaguier A, Kochko Ad, Dubreuil-Tranchant C, This P, Boursiquot J-M, Le Cunff L. 2011. An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *American Journal of Botany* **98**(1): e13-e15.

Chabin JP, Madelin M, Bonnefoy C. 2008. Réchauffement climatique, quels impacts probables sur les vignobles ? *Université de Bourgogne – Centre de climatologie Article*.

Champagnol F. 1984. Elements de physiologie de la vigne et viticulture générale. *Ed Dehan,Montpellier, France*.

Chen J, Chuzhanova N, Stenson PD, Férec C, Cooper DN. 2005. Complex gene rearrangements caused by serial replication slippage. *Human Mutation* **26**(2): 125-134.

Chen N. 2002. *Using repeatmasker to identify repetitive elements in genomic sequences*.

Cheng X, Zhang D, Cheng Z, Keller B, Ling HQ. 2009. A new family of ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. *Genetics* **181**(4): 1183-1193.

Ching A, Caldwell K, Jung M, Dolan M, Smith O, Tingey S, Morgante M, Rafalski A. 2002. Snp frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *Bmc Genetics* **3:19**.

Cloutier D, Rioux D, Beaulieu J, Schoen DJ. 2003. Somatic stability of microsatellite loci in eastern white pine, pinus strobus l. *Heredity* **90**(3): 247-252.

Crow JF. 1992. An advantage of sexual reproduction in a rapidly changing environment. *Journal of Heredity* **83**(3): 169-173.

Darwin C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.

DeBarry J, Liu R, Bennetzen J. 2008. Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the assisted automated assembler of repeat families (aaarf) algorithm. *BMC Bioinformatics* **9**(1): 235.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**(5): 491-498.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in arabidopsis. *Genome Research* **12**(7): 1075-1079.

Diakou P, Moing A, Svanella L, Ollat N, Rolin DB, Gaudillere M, Gaudillere JP. 1997. Biochemical comparison of two grape varieties differing in juice acidity. *Australian Journal of Grape and Wine Research* **3**(3): 1-10.

Dion R. 1982. Histoire de la vigne et du vin en france des origines au 19éme siécle. *Ed. Flammarion: Paris* **XII**.

Doerfler W, Böhm P, Haaf T. 2006. Methylation dynamics in the early mammalian embryo: Implications of genome reprogramming defects for development. *DNA methylation: Development, genetic disease and cancer*: Springer Berlin Heidelberg, 13-22.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*.

Doligez A, Adam-Blondon A, Cipriani G, Di Gaspero G, Laucou V, Merdinoglu D, Meredith C, Riaz S, Roux C, This P. 2006. An integrated ssr map of grapevine based on five mapping populations. *TAG Theoretical and Applied Genetics* **113**(3): 369-382.

Dong C, Whitford R, Langridge P. 2002. A DNA mismatch repair gene links to the ph2 locus in wheat. *Genome* **45**(1): 116-124.

Doré C, Varoquaux F. 2006. Histoire et amélioration de cinquante plantes cultivées *Quae Edition*.

Duchene E, Legras JL, Karst F, Merdinoglu D, Claudel P, Jaegli N, Pelsy F. 2009. Variation of linalool and geraniol content within two pairs of aromatic and non-aromatic grapevine clones. *Australian Journal of Grape and Wine Research* **15**(2): 120-130.

Edwards A, Caskey CT. 1991. Closure strategies for random DNA sequencing. *Methods* **3**: 41 - 47.

Eichler EE, Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**(5634): 793-797.

Einset J, Pratt C. 1954. 'giant' sports of grapes. *Proc Am Soc Horticult Sci* **63**(251-256).

Emanuelli F, Battilana J, Costantini L, Le Cunff L, Boursiquot J, M, This P, Grando M. 2010. A candidate gene association study on muscat flavor in grapevine (vitis vinifera l.). *BMC Plant Biology* **10**(1): 241.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces usingphred ii. Error probabilities. *Genome Research* **8**(3): 186-194.

Falavigna A, Casali PE. 2002. Pratical aspects of a breeding program of asparagus based on in vitro anther culture. *Acta Hort.* **583**: 201-210.

Fan F, Liu Cp, Tavare S, Arnheim N. 1999. Polymorphisms in the human DNA repair gene xpf. *Mutation Research/Mutation Research Genomics* **406**(2): 115-120.

Fernandez L, Doligez A, Lopez G, Thomas MR, Bouquet A, Torregrosa L. 2006. Somatic chimerism, genetic inheritance, and mapping of the fleshless berry (flb) mutation in grapevine (vitis vinifera l). *Genome* **49**: 721-728.

Fernandez L, Torregrosa L, Segura V, Bouquet A, Martinez-Zapater JM. 2010. Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. *The Plant Journal* **61**(4): 545-557.

Fernandez L, Torregrosa L, Terrier N, Sreekantan L, Grimplet J, Davies C, Thomas M, Romieu C, Ageorges A. 2007. Identification of genes associated with flesh morphogenesis during grapevine fruit development. *Plant Molecular Biology* **63**(3): 307-323.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**(5): 397-405.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: Where genetics meets genomics. *Nat Rev Genet* **3**: 329 - 341.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat. Rev. Genet.* **7**: 85-97.

Flicek P, Birney E. 2009. Sense from sequence reads: Methods for alignment and assembly. *Nat Meth* **6**(11s): S6-S12.

Fournier-Level A, Lacombe T, Le Cunff L, Boursiquot JM, This P. 2010. Evolution of the vvmyba gene family, the major determinant of berry colour in cultivated grapevine (vitis vinifera l.). *Heredity* **104**(4): 351-362.

Fournier-Level A, Le Cunff L, Gomez C, Doligez A, Ageorges A, Roux C, Bertrand Y, Souquet JM, Cheynier V, This P. 2009. Quantitative genetic bases of anthocyanin variation in grape (vitis vinifera l. Ssp sativa) berry: A qtl to qtn integrated study. *Genetics* **183**: 1127 - 1139.

Franks T, Botta R, Thomas MR, Franks J. 2002. Chimerism in grapevines: Implications for cultivar identity, ancestry and genetic improvement. *TAG Theoretical and Applied Genetics* **104**(2): 192-199.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C. 2006. Copy number variation: New insights in genome diversity. *Genome Research* **16**(8): 949-961.

Gaffiot F. 1934. Dictionnaire latin français. *Hachette Livre*.

Galet P. 1985. Précis d'ampélographie pratique *lavoisier Livre*.

Galet P. 1988. Cépages et vignobles de france *Pierre GALET*.

Galet P. 1993. Précis de viticulture. *Ed. Galet*.

Galet P. 2000. Dictionnaire encyclopedique des cépages. *Hachette book*.

Giannuzzi G, D'Addabbo P, Gasparro M, Martinelli M, Carelli F, Antonacci D, Ventura M. 2010. Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics* **12**(1): 436.

Gill DE, Chao L, Perkins SL, Wolf JB. 1995. Genetic mosaicism in plants and clonal animals. *Annual Review of Ecology and Systematics* **26**(1): 423-444.

Godoy VG, Fox MS. 2000. Transposon stability and a role for conjugational transfer in adaptive mutability. *Proceedings of the National Academy of Sciences* **97**(13): 7393-7398.

Gore MA, Wright MH, Ersoz ES, Bouffard P, Szekeres ES, Jarvie TP, Hurwitz BL, Narechania A, Harkins TT, Grills GS, Ware DH, Buckler ES. 2009. Large-scale discovery of gene-enriched snps. *The Plant Genome* **2**(2): 121-133.

Grandbastien MA. 1998. Activation of plant retrotransposons under stress conditions. *Trends in Plant Science* **3**(5): 181-187.

Grapputo A, Kumpulainen T, Mappes J, Parri S. 2005. Genetic diversity in populations of asexual and sexual bag worm moths (lepidoptera: Psychidae). *BMC Ecology* **5**(1): 5.

Gray YH. 2000. It takes two transposons to tango:Transposable-element-mediated chromosomal rearrangements. *Trends in Genetics* **16**(10): 461-468.

Green RE, Briggs AW, Krause J, Prufer K, Burbano HA, Siebauer M, Lachmann M, Paabo S. 2009. The neandertal genome and ancient DNA authenticity. *EMBO J* **28**(17): 2494-2502.

Grenan S, Bonnet A, Boidron R. 2000. Result and thoughts on 35 years of sanitary selection in france *Acta Hort.* **528**(713-722).

Grenan S, Truel P. 1983. Observations sur un aspect de la variabilité constatée au cours de la multiplication végétative de variétés de vigne issues de semis de vitis vinifera l. *Agronomie* **3**(7): 675-680.

Gustafsson A. 1947. Mutations in agricultural plants. *Hereditas* **33**(1-2): 1-100.

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC. 2011. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26**(10): 1277-1283.

Harismendy O, Ng P, Strausberg R, Wang X, Stockwell T, Beeson K, Schork N, Murray S, Topol E, Levy S, Frazer K. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**(3): R32.

Hartmann HT, Kester DE, Davis Jr FT, Geneve RL 2001. Plant propagation: Principles and practices.In: The Genetics Society.

Hashida SN, Kitamura K, Mikami T, Kishima Y. 2003. Temperature shift coordinately changes the activity and the methylation state of transposon tam3 in antirrhinum majus. *Plant Physiology* **132**(3): 1207-1216.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**(2): 1157-1164.

Hedges D, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Zachner S. 2009. Exome sequencing of a multigenerational human pedigree. *PLos One* **4**(12): e8232.

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. 1996. Retrotransposons of rice involved in mutations induced by tissue culture. *Proceedings of the National Academy of Sciences of the United States of America* **93**(15): 7783-7788.

Hocquigny S, Pelsy F, Dumas V, Kindt S, Heloir MC, Merdinoglu D. 2004. Diversification within grapevine cultivars goes through chimeric states. *Genome* **47**: 579-589.

Holliday R. 2006. Epigenetics: A historical overview. *Epigenetics* **1**(2): 76-80.

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. 2010. Next-generation variationhunter: Combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**(12): i350-i357.

Huang W, Marth G. 2008. Eagleview: A genome assembly viewer for next-generation sequencing technologies. *Genome Research* **18**(9): 1538-1543.

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. *Trends in Genetics* **24**(5): 238-245.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**(9): 949-951.

Imazio S, Labra M, Grassi F, Winfield M, Bardini M, Scienza A. 2002. Molecular tools for clone identification: The case of the grapevine cultivar “traminer”. *Plant Breeding* **121**: 531-535.

Inoue K, Lupski JR. 2002. Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics* **3**(1): 199-242.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463 - 467.

Jiang J, Birchler JA, Parrott WA, Dawe RK. 2003. A molecular view of plant centromeres. *Trends in Plant Science* **8**(12): 570-575.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet* **9**: 418-420.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**(1-4): 462-467.

Kaiser J. 2008. A plan to capture human diversity in 1000 genomes. *Science* **319**(5862): 395.

Kankel MW, Ramsey DE, Stokes TL, Flowers SK, Haag JR, Jeddeloh JA, Riddle NC, Verbsky ML, Richards EJ. 2003. Arabidopsis met1 cytosine methyltransferase mutants. *Genetics* **163**(3): 1109-1122.

Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences* **98**(15): 8714-8719.

Kazazian HH. 2004. Mobile elements: Drivers of genome evolution. *Science* **303**(5664): 1626-1632.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49 - 63.

Klekowski EJ. 1998. Mutation rates in mangroves and other plants. *Genetica* **102-103**(0): 325-331.

Klekowski EJ, Godfrey PJ. 1989. Ageing and mutation in plants. *Nature* **340**: 389-391.

Knox M, Moreau C, Lipscombe J, Baker D, Ellis N 2009. High-throughput retrotransposon-based fluorescent markers: Improved information content and allele discrimination.In: Plant Methods.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science* **304**(5673): 982.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2005. Association of vvmyba1 gene expression with anthocyanin production in grape (vitis vinifera) skin-color mutants. *Journal of the Japanese Society for Horticultural Science* **74**(3): 196-203.

Kobayashi S, Ishimaru M, Hiraoka K, Honda C. 2002. Myb-related genes of the kyoho grape (vitis labruscana) regulate anthocyanin biosynthesis. *Planta* **215**: 924 - 933.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An snp caused loss of seed shattering during rice domestication. *Science* **312**(5778): 1392-1396.

Konradi J, Blaich R, Forneck A. 2007. Genetic variation among clones and sports of "pinot noir" (vitis vinifera l.). *Eur. J Horticulture Science* **72**(6): 275-279.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**(5849): 420-426.

Labate J, Robertson L, Wu F, Tanksley S, Baldo A. 2009. Est, cosii, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (solanum lycopersicum l.). *Theor Appl Genet*(118): 1005-1014.

Labra M, Imazio S, Grassi F, Rossoni M, Sala F. 2004. Vine-1 retrotransposon-based sequence-specific amplified polymorphism for vitis vinifera l. Genotyping. *Plant Breeding* **123**(2): 180-185.

Lacombe T. 2009. Historique du domaine de vassal. *http://www1.montpellier.inra.fr/vassal/index.html*.

Lacombe T, Laucou V, Vecchi-Staraz M, This P, Boursiquot JM. In prep. Large-scale parentage analysis in an extended set of grapevine cultivars (vitisvinifera).

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**(3): 1-10.

Laucou V, Lacombe T, Dechesne F, Siret R, Bruno JP, Dessup M, Dessup T, Ortigosa P, Parra P, Roux C, Santoni S, Varès D, Péros JP, Boursiquot JM, This P. 2011. High throughput analysis of grape genetic diversity as a tool for germplasm collection management. *TAG Theoretical and Applied Genetics* **122**(6): 1233-1245.

Lauria M, Rossi V. 2011. Epigenetic control of gene regulation in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **In Press, Accepted Manuscript**.

Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon AF, Boursiquot JM, This P. 2008. Construction of nested genetic core collections to optimize the exploitation of natural diversity in vitis vinifera l. Subsp sativa. *BMC Plant Biology* **8**.

Ledergerber C, Dessimoz C. 2010. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*.

Legros JP. 1993. L'invasion du vignoble par le phylloxéra. *Academie des sciences et lettres de Montpellier* **Bull. n°24**(205-22).

Legros JP. 1997. Le phylloxéra, une histoire sans fin. *Association française pour l'avancement des sciences* **97**(1): 32-39.

Levadoux L. 1956. Les populations sauvages et cultivées de vitis vinifera l. *Ann Amélioration des plantes* **6**: 59-118.

Levy AA, Feldman M. 2002. The impact of polyploidy on grass genome evolution. *Plant Physiology* **130**(4): 1587-1593.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**: 1754 - 1760.

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. *Bioinformatics* **25**(16): 2078-2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**(11): 1851-1858.

Li R, Li Y, Kristiansen K, Wang J. 2008. Soap: Short oligonucleotide alignment program. *Bioinformatics* **24**(5): 713-714.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. 2009. Soap2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**(15): 1966-1967.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Molecular Ecology* **11**(12): 2453-2465.

Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM. 2007. High throughput snp discovery and genotyping in grapevine (vitis vinifera l.) by combining a re-sequencing approach and snplex technology. *BMC Genomics* **8**: 424 - 434.

Linnarsson S. 2010. Recent advances in DNA sequencing methods, general principles of sample preparation. *Experimental Cell Research* **316**(8): 1339-1343.

Llorens C, Munoz-Pomer A, Bernad L, Botella H, Moya A. 2009. Network dynamics of eukaryotic ltr retroelements beyond phylogenetic trees. *Biology Direct* **4**(1): 41.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat Rev Genet* **4**(11): 865-875.

Longley MJ, Graziewicz MA, Bienstock RJ, Copeland WC. 2005. Consequences of mutations in human DNA polymerase î³. *Gene* **354**(0): 125-131.

Mannini F, Argamante N, Credi R. 1999. Contribution of virus infections to clonal variability of some vitis vinifera l. Cultivars. *Bulletin de l'OIV* **72**(817-818).

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**(3): 133-141.

Margraf RL, Durtschi JD, Dames S, Pattison DC, Stephens JE, Mao R, Voelkerding KV. 2010. Multi-sample pooling and illumina genome analyzer sequencing methods to determine gene sequence variation for database development. *Journal of biomolecular techniques : JBT* **21**(3): 126-140.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376 - 380.

Martinez MC, Boursiquot JM, Grenan S, Boidron R. 1997. Etude ampélométrique de feuilles adultes de somaclones du cv. Grenache n (vitis vinifera l.). *Canadian Journal of Botany* **75**: 333-345.

Mc Govern PE. 2003. Ancient wine: The search for the origins of viniculture. *Princeton: Princeton University Press*.

McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, Duerr RH, Silverberg MS, Taylor KD, Rioux JD, Altshuler D, Daly MJ, Xavier RJ. 2008. Deletion polymorphism upstream of irgm associated with altered irgm expression and crohn's disease. *Nat Genet* **40**(9): 1107-1112.

McClintock B. 1953. Induction of instability at selected loci in maize. *Genetics* **38**(6): 579-599.

McClintock B. 1984. Significance of responses of the genome to challenge. *Science* **226**: 792-801.

McDonald JF. 1995. Transposable elements: Possible catalysts of organismic evolution. *Trends in Ecology & Evolution* **10**(3): 123-126.

McIntyre CL, Drenth J, Gonzalez N, Henzell RG, Jordan DR. 2008. Molecular characterization of the waxy locus in sorghum. *Genome* **51**(7): 524-533.

McKey D, Elias M, Pujol B, Duputié A. 2009. The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist* **186**(2): 318-332.

McNally KL, Bruskiewich R, Mackill D, Buell CR, Leach JE, Leung H. 2006. Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. *Plant Physiol* **141**(1): 26-31.

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat Meth* **6**(11s): S13-S20.

Mefford HC, Clauin S, Sharp AJ, Moller RS, Ullmann R, Kapur R, Pinkel D, Cooper GM, Ventura M, Ropers HH, Tommerup N, Eichler EE, Bellanne-Chantelot C. 2007. Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *The American Journal of Human Genetics* **81**(5): 1057-1069.

Meneghetti S, Costacurta A, Frare E, Da Rold G, Migliaro D, Morreale G, Crespan M, Sotés V, Calò A. 2011. Clones identification and genetic characterization of garnacha grapevine by means of different pcr-derived marker systems. *Molecular Biotechnology* **48**(3): 244-254.

Metzker ML. 2010. Sequencing technologies [mdash] the next generation. *Nat Rev Genet* **11**(1): 31-46.

Meyer E, Aglyamova G, Wang S, Buchanan-Carter J, Abrego D, Colbourne J, Willis B, Matz M. 2009. Sequencing and de novo analysis of a coral larval transcriptome using 454 gsflx. *BMC Genomics* **10**(1): 219.

Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O. 2009. Selective epigenetic control of retrotransposition in arabidopsis. *Nature* **461**(7262): 427-430.

Moisy C. 2008. Analyse structurale et transciptionnelle des rétrontransposons du génome de la vigne. *Doctorat de l'Université Louis Paster de Strasbourg*.

Moisy C, Blanc S, Merdinoglu D, Pelsy F. 2008. Structural variability of tvv1 grapevine retrotransposons can be caused by illegitimate recombination. *Theor Appl Genet* **116**: 671 - 682.

Moisy C, Garrison K, Meredith C, Pelsy F. 2008. Characterization of ten novel ty1/copia-like retrotransposon families of the grapevine genome. *BMC Genomics* **9**(1): 469.

Moncada X, Hinrichsen R P. 2007. Limited genetic diversity among clones of red wine cultivar "Carmenère" As revealed by microsatellite an aflp markers *Vitis* **46**(4): 174-181.

Moncada X, Pelsy F, Merdinoglu D, Hinrichsen P. 2006. Genetic diversity and geographical dispersal in grapevine clones revealed by microsatellite markers. *Genome* **49**: 1459-1472.

Morales-Ruiz T, Ortega-Galisteo AP, Ponferrada-Marin MI, Martinez-Macias MI, Ariza RR, Roldan-Arjona T. 2006. Demeter and repressor of silencing 1 encode 5-methylcytosine DNA glycosylases. *Proceedings of the National Academy of Sciences* **103**(18): 6853-6858.

Morgan HD, Sutherland HGE, Martin DIK, Whitelaw E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics* **23**: 314-318.

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, Bustamante CD, Buckler ES. 2011. Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences* **108**(9): 3530-3535.

Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D. 2010. Rapid genomic characterization of the genus vitis. *PLos One* **5**(1): e8219.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1): 297-304.

Nagy Z, Chandler M. 2004. Regulation of transposition in bacteria. *Research in Microbiology* **155**(5): 387-398.

Neilson Jones W. 1969. Plant chimeras. 2nd edn. *Methuen :London*.

Neumann P, Pozarkova D, Macas J. 2003. Highly abundant pea ltr retrotransposon ogre is constitutively transcribed and partially spliced. *Plant Mol Biol* **53**(3): 399-410.

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic variation in an individual human exome. *PLoS Genet* **4**(8): e1000160.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and snp calling from next-generation sequencing data. *Nat Rev Genet* **12**(6): 443-451.

Nordborg M, Weigel D. 2008. Next-generation genetics in plants. *Nature* **456**(7223): 720-723.

O'Hare K, Rubin GM. 1983. Structures of p transposable elements and their sites of insertion and excision in the drosophila melanogaster genome. *Cell* **34**(1): 25-35.

OIV. 2007. *2ème édition de la liste des descripteurs oiv pour la description des variétés et espèces de vitis*. Paris.

Olmo HP. 1976. Grapes. In:Evolution of crop plants. *N.W Simmonds, Longman, London*.

Orive ME. 2001. Somatic mutations in organisms with complex life histories. *Theoretical Population Biology* **59**(3): 235-249.

Pardue ML, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL. 2005. Two retrotransposons maintain telomeres in drosophyla. *Chromosome Research* **13**(5): 443-453.

Park SM, Hiramatsu M, Wakana A. 1999. Aneuploid plants derived from crosses with triploid grapes through immature seed culture and subsequent embryo culture. *Plant Cell* **59**: 125-133.

Pelsy F. 2009. Molecular and cellular mechanisms of diversity within grapevine varieties. *Heredity* **104**: 331-340.

Pelsy F, Hocquigny S, Moncada X, Barbeau G, Forget D, Hinrichsen P, Merdinoglu D. 2010. An extensive study of the genetic diversity within seven french wine grape variety collections. *TAG Theoretical and Applied Genetics* **120**(6): 1219-1231.

Pelsy F, Merdinoglu D. 2002. Complete sequence of tvv1, a family of ty1 copia-like retrotransposons of vitis vinifera l., reconstituted by chromosome walking. *Theor Appl Genet* **105**: 614 - 621.

Pereira H, Barao A, Delgado M, Morais-Cecílio L, Viegas W. 2005. Genomic analysis of grapevine retrotransposon gret1 in vitis vinifera. *TAG Theoretical and Applied Genetics* **111**(5): 871-878.

Péros JP, Berger G, Portemont A, Boursiquot JM, Lacombe T. 2010. Genetic variation and biogeography of the disjunct vitis subg. Vitis (vitaceae). *Journal of Biogeography* **38**(3): 471-486.

Perrier X, Jacquemond Collet J. 2006. Darwin software. *http://darwin.cirad.fr/darwin/*

Petit RJ, Hampe A. 2006. Some evolutionary consequences of being a tree. *Annual Review of Ecology, Evolution, and Systematics* **37**(1): 187-214.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in oryza australiensis, a wild relative of rice. *Genome Research* **16**(10): 1262-1269.

Piperi C, Papavassiliou AG. 2011. Strategies for DNA methylation analysis in developmental studies. *Development, Growth & Differentiation* **53**(3): 287-299.

Pouget R. 1990. Histoire de la lutte contre le phylloxéra de la vigne en france : 1868-1895. *Institut national de la recherche agronomique*.

Ramensky V, Bork P, Sunyaev S. 2002. Human non synonymous snps: Server and survey. *Nucleic Acids Research* **30**(17): 3894-3900.

Rapp RA, Wendel JF. 2005. Epigenetics and plant evolution. *New Phytologist* **168**(1): 81-91.

Regner F, Stadlhuber C, Eisenheld C, Kaserer H. 2000. Considerations about the evolution of grapevine and the role of traminer. *Acta Horticulturae* **528**.

Riaz S, Garisson KE, Dangl GS, Boursiquot JM, Meredith CP. 2002. Genetic divergence and chimerism within ancient asexually propagated winegrape cultivars. *J Am Soc Hortic Sci* **127**: 508-514.

Rice WR. 2002. Experimental tests of the adaptive significance of sexual recombination. *Nat Rev Genet* **3**(4): 241-251.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotech* **29**(1): 24-26.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242**(1): 84-89.

Royer C. 1988. Mouvement historiques de la vigne dans le monde. In la vigne et le vin *La Manufacture et la Cité des sciences et de l'industrie*: 15–25.

Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegu B, Roulin A, Guiderdoni E, Delabastide M, McCombie R, Panaud O. 2011. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. *The Plant Journal* **66**(2): 241-246.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406 - 425.

Sakai H, Tanaka T, Itoh T. 2007. Birth and death of genes promoted by transposable elements in oryza sativa. *Gene* **392**(1-2): 59-63.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The b73 maize genome: Complexity, diversity, and dynamics. *Science* **326**(5956): 1112-1115.

Schön I, Martens K, Dijk P, Forneck A, Benjak A, Rühl E 2009. Grapevine (vitis ssp): Example of clonal reproduction in agricultural important plants. *Lost sex*: Springer Netherlands, 581-598.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin Pr, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**(5683): 525-528.

Seberg O, Petersen G. 2009. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat Rev Genet* **10**(4): 276-276.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJL, Eichler EE. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38**(9): 1038-1042.

Sharp AJ, Selzer RR, Veltman JA, Gimelli S, Gimelli G, Striano P, Coppola A, Regan R, Price SM, Knoers NV, Eis PS, Brunner HG, Hennekam RC, Knight SJL, de Vries BBA, Zuffardi O, Eichler EE. 2007. Characterization of a recurrent 15q24 microdeletion syndrome. *Human Molecular Genetics* **16**(5): 567-572.

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research* **10**(7): 908-915.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**(4): 272-285.

Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijó JA, Martienssen RA. 2009. Epigenetic reprogramming and small rna silencing of transposable elements in pollen. *Cell* **136**(3): 461-472.

Smit, A. H, R & Green. 1996-2004. Repeatmasker open-3.0. http://www.repeatmasker.org.

Spilmont AS. 2005. Déperissement de la syrah, compte rendu de la réunion du groupe de travail, 11 avril 2005. *Prog. Agr. Vit.* **122**: 15-16.

Spilmont AS, Moreno Y, Audeguin L. 2010. Dépérissement de la syrah : Des symptômes similaires sur des franc-de-pied. *Prog. Agric. Vitic.* **3**(63-67).

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, TavarÃ© S, Deloukas P, Hurles ME, Dermitzakis ET. 2007a. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**(5813): 848-853.

Sunyaev S, Kondrashov FA, Bork P, Ramensky V. 2003. Impact of selection, mutation rate and genetic drift on human genetic variation. *Human Molecular Genetics* **12**(24): 3325-3330.

Taylor KH, Kramer RS, Davis JW, Guo J, Duff DJ, Xu D, Caldwell CW, Shi H. 2007. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Research* **67**(18): 8511-8518.

The Angiosperm Phylogeny Group. 2009. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: Apg iii. *Botanical Journal of the Linnean Society* **161**(2): 105-121.

This P, Jung A, Boccacci P, Borrego J, Botta R, Costantini L, Crespan M, Dangl G, Eisenheld C, Ferreira-Monteiro F, Grando S, Ibáñez J, Lacombe T, Laucou V, Magalhães R, Meredith C, Milani N, Peterlunger E, Regner F, Zulini L, Maul E. 2004. Development of a standard set of microsatellite reference alleles for identification of grape cultivars. *TAG Theoretical and Applied Genetics* **109**(7): 1448-1458.

This P, Lacombe T, Cadle-Davidson M, Owens CL. 2007. Wine grape (vitis vinifera l.) color associates with allelic variation in the domestication gene vvmyba1. *Theor Appl Genet* **114**(4): 723 - 730.

This P, Lacombe T, Thomas MR. 2006. Historical origins and genetic diversity of wine grapes. *Trends in Genetics* **22**(9): 511-519.

Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**(22): 4673-4680.

Thompson M, Olmo H. 1963. Cytohistological studies of cytochimeric and tetraploid grapes. *Am J Botany* **50**: 901-906.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* **28**: 286 - 289.

Troggio M, Malacarne G, Coppola G, Segala C, Cartwright DA, Pindo M, Stefanini M, Mank R, Moroldo M, Morgante M, Grando MS, Velasco R. 2007. A dense single-nucleotide polymorphism-based genetic linkage map of grapevine (vitis vinifera l.) anchoring pinot noir bacterial artificial chromosome contigs. *Genetics* **176**(4): 2637-2650.

Valleau WD. 1916. Inheritance of sex in the grape. *The American Naturalist*.

Van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Research* **15**(9): 1243-1249.

Van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics* **19**(10): 530-536.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Dematta L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLos One* **2**(12): e1326.

Verries C, Bes C, This P, Tesniere C. 2000. Cloning and characterization of vine-1, a ltr- retrotransposon-like element in vitis vinifera l and other vitis species. *Genome* **43**: 366-376.

Vezzulli S, Troggio M, Coppola G, Jermakow A, Cartwright D, Zharkikh A, Stefanini M, Grando M, Viola R, Adam-Blondon AF, Thomas M, This P, Velasco R. 2008. A reference integrated map for cultivated grapevine (vitis vinifera l.) from three crosses, based on 283 ssr and 501 snp-based markers. *TAG Theoretical and Applied Genetics* **117**(4): 499-511.

Viala P, Vermorel V. 1910. Traité général d'ampélographie. *Ed. Masson, Paris*.

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J. 2002. Rate and pattern of mutation at microsatellite loci in maize. *Mol Biol Evol* **19**(8): 1251-1260.

Wagner, Antcliff. 1980. A study of sexual progenies of bicane x sultanina (vitis vinifera l.), evidence for genetic differences between sultana clones in berry weight. *Dept. of Vitic. and EnologyDavis, Calif. (USA)*.

Walker AR, Lee E, Robinson SP. 2006. Two new grape cultivars, bud sports of cabernet sauvignon bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry colour locus. *Plant Mol Biol* **62**(4-5): 623-635.

Walter B, Martelli G. 1998. Considerations on grapevineselection and certification. *Vitis* **37**(87-90).

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, san A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK-S, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. 2008. The diploid genome sequence of an asian individual. *Nature* **456**(7218): 60-65.

Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BBT, Powell W. 1997. Genetic distribution of bare–1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (s-sap). *Molecular and General Genetics MGG* **253**(6): 687-694.

Wegscheider E, Benjak A, Forneck A. 2009. Clonal variation in pinot noir revealed by s-sap involving universal retrotransposon-based sequences. *Am. J. Enol. Vitic.* **60**(1): 104-109.

Weigel D, Mott R. 2009. The 1001 genomes project for arabidopsis thaliana. *Genome Biology* **10**(5): 107.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**(12): 973-982.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2009. Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat Rev Genet* **10**(4): 276-276.

Wu JL, Wu C, Lei C, Baraoidan M, Bordeos A, Madamba M, Ramos-Pamplona M, Mauleon R, Portugal A, Ulat V, Bruskiewich R, Wang G, Leach J, Khush G, Leung H. 2005. Chemical- and irradiation-induced mutants of indica rice ir64 for forward and reverse genetics. *Plant Molecular Biology* **59**(1): 85-97.

Yamashita H, Shigehara I, Haniuda T. 1998. Production of triploid grapes by in ovulo embryo culture. *Vitis* **37**(3): 113-117.

Yamashita S, Takano-Shimizu T, Kitamura K, Mikami T, Kishima Y. 1999. Resistance to gap repair of the transposon tam3 in antirrhinum majus: A role of the end regions. *Genetics* **153**(4): 1899-1908.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF, Hoover R, Hunter DJ, Chanock SJ,

Thomas G. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**(5): 645-649.

Yobregat O, Sereno C, Audeguin L, Lacombe T, Boursiquot JM. 2011. Conservation de la diversité intravariétale de la vigne : Situation générale en 2010. Perspectives et priorités pour l'avenir. *Progrès agricole et viticole* **10**.

Zhang Z, Lin H, Ma B. 2010. Zoom lite: Next-generation sequencing data mapping and visualization software. *Nucleic Acids Research* **38**(suppl 2): W743-W748.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB. 2003. Single nucleotide polymorphisms in soybean. *Genetics* **163**: 1123-1134.

Zienolddiny S, Campa D, Lind H, Ryberg D, Skaug V, Stangeland L, Phillips DH, Canzian F, Haugen A. 2006. Polymorphisms of DNA repair genes and risk of non-small cell lung cancer. *Carcinogenesis* **27**(3): 560-567.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39**: 61-69.

Annexes

Annexe-1 : Protocole d'extraction d'ADN pour la nouvelle génération de séquenceur

Detailed protocol for plant nuclear DNA preparation for next sequencing methods

PROCEDURE

NUCLEI EXTRACTION

1. Quickly grind 5 to 6 g (fresh weight) of plant material in liquid nitrogen to a fine powder using a mortar and pestle. Transfer the fine powder into a 50 mL tube before cooling in liquid nitrogen.

 All the following steps of the nuclei extraction should be done under a biological hood.

2. Quickly add 45 mL of SEBM into the 50 mL tube.
3. Incubate the sample on ice for 12 minutes, mixing regularly.
4. Filter the sample through Miracloth twice.
5. Filter 15 mL of the sample with a cell strainer (d: 40 µm) into new 50 mL tubes (3 tubes in total).
6. Add Triton X-100 (10 %) to each tube and incubate the sample on ice for 12 minutes, mixing regularly.
7. Centrifuge at 600 g for 9 minutes at 4 °C. Discard the supernatant.
8. Add 20 mL of SEB to each tube and mix to resuspend the pellet.
9. Filter 60 ml obtained at the step 8 (3 X 20 ml) into 2 new 50 mL tubes with the cell strainer.
10. Centrifuge at 600 g for 9 minutes at 4 °C. Discard the supernatant.
11. Add 20 mL of SEB and mix to resuspend the pellet.
12. Filter the content of the 2 tubes (20 + 20 = 40 mL) into a new 50 mL tube through a cell strainer.
13. Centrifuge at 600 g for 9 minutes at 4 °C. Discard the supernatant.
14. Add 2 mL of SEB and mix to resuspend the pellet. Store the suspension of the purified nuclei at 4 °C.

NUCLEAR DNA PURIFICATION

1. Add 4 mL of NLB to the suspension of purified nuclei. Incubate at 55 °C for 3 hours with mixing.
2. Add 4 mL AcK (5M/3M) and mix for 10 minutes.
3. Centrifuge at 3000 g for 10 minutes at 4 °C.
4. Transfer the supernatant into a 50 mL tube. Add 15 mL of DBB and 250 µL of silica matrix. Mix by inverting the tube for 10 minutes and then centrifuge at 1500 g for 2 minutes at 4 °C.
5. Discard the supernatant, add 5 mL of DWS and mix for 2 minutes to resuspend the pellet.
6. Centrifuge at 1500 g for 2 minutes at 4 °C. Repeat this step one more time.
7. Discard the supernatant and place the tube, without its cap, at 37 °C for 10 minutes.
8. Add 600 µL water and incubate at 37 °C for 10 minutes.
9. Transfer the sample into 2 mL microtube. Centrifuge at 14 000 g for 3 minutes at room temperature
10. Carefully transfer the supernatant into a new microtube by pipeting.
11. Add 60 µL of 3 M Na-Ac and 1 mL EtOH at 96 %. Mix for 2 minutes and place the tube at -20°C for 5 to 10 minutes.
12. Centrifuge at 14 000 g for 5 minutes. Discard the supernatant and wash the DNA with 250µL EtOH 75%.
13. Centrifuge at 14 000g for 5 minutes.
14. Discard supernatant and add 120 µL of sterile ultra-pure water. Store the DNA at -20 °C until use.

RECIPES

Nucleus Extraction Buffer
Abbreviation: NEB; 1 L for ten samples

Tris	100mM (pH=8)
KCl	1M
EDTA	100 mM

Storage 4 °C

Sucrose-Based Extraction Buffer
Abbreviation: SEB; 500 mL for 2 samples

NEB	10 %
Sucrose	550 mM
Spermidinetrihydrochloride	4 mM
SpermidineTetrahydrochloride	1 mM
carbamic acid	0.13 % (p/v)
PVP40	0.25 % (p/v)

No storage

Sucrose-Based Extraction Buffer with 2-Mercaptoethanol
Abbreviation: SEBM; 250 mL for 2 samples

SEB	250 mL
2-Mercaptoethanol	0.2% (v/v)

No storage

NuclearLysis Buffer
Abbreviation: NLB; 100 mL for 2 samples

EDTA	400 mM
N Lauryl sarkosyl	2% (p/v)
Proteinase K	1mg/ml

Storage 4 °C

DNA Binding Buffer
Abbreviation: DBB; 50 mL for many samples
Chloride of guanidium260 mM in Ethanol 96%
Room temperature

DNA Wash Solution
Abbreviation: DWS; 100 mL for many samples

Tris (pH=8)	22.5 mM
AcK	160 mM
EDTA	1 mM
Ethanol 96%	(1.7:1), For a total 100mL, add 170ml Ethanol 96%

Room temperature

Acetate of K
Abbreviation:AcK; 100 mL for many samples

Aceticacid	11.5%
Potassium acetate	5M

Room temperature

Silica matrix preparation

5 g of silicon dioxide (Sigma, S5631) is mixed with 50 mL of ultra-pure water.
Settled for 30 minutes at room temperature, then centrifuged at 700g for 3 minutes.
The supernatant is removed and the pellet resuspended with 40 mL of ultra-pure water, then re-settled for another 10 minutes.
The supernatant is removed and the pellet resuspended with 5 mL of ultra-pure water, then re-settled for another 10 minutes.
The tube is centrifuged at 700g for 3 minutes, the supernatant is removed, and the pellet then resuspended with 1.7 mL of ultra-pure water and mixed vigorously.
The total volume of the silica matrix suspension is approximately 5 mL; 50 µl of HCl (37%, wt/vol) is added to adjust the suspension to pH 2.
Small aliquotof 0.5 ml are transferred into 2 ml Eppendorf tubes and autoclaved for 20 min at 121°C. The silica matrix suspension at a final concentration of 1 g/ml remains stable for at least 12 months when stored at 4°C.

REAGENTS

Product	chemical **Formula**	**Vendor**	**Reference**
EDTA (EthylenediaminetetraaceticAcid)	$C_{10}H_{14}N_2Na_2O_8 + 2H_20$	Sigma	139-33-3
Sodium hydroxide	NaOH	Sigma	1310-73-2
Tris (Trisma Base, Tris[hydroxymethyl]aminometane)	$C_4H_{11}NO_3$	Sigma	77-86-1
Hydrochloricacid	HCl	Sigma	7647-01-0
Potassium chloride	KCl	Sigma	7447-40-7
Sucrose	$C_{12}H_{22}O_{11}$	Sigma	57-50-1
Spermidinetrihydrochloride	$NH_2(CH_2)_3NH(CH_2)_4NH_2 + 3HCl$	Sigma	334-50-9
Spermine tetrahydrochloride	$C_{10}H_{26}N_4 + 4HCl$	Sigma	306-67-2
“carbamic acid”, Sodium diethyldithiocarbamatetrihydrate	$(C_2H_5)_2NCSSNa + 3H_20$	Sigma	20624-25-3
Polyvinylpyrrolidone PVP (PVP 40000, PVT 40T)	$[C_6H_9NO]_n$	Sigma	9003-39-8
2-mercaptoethanol (beta-mercaptoethanol)	$HSCH_2CH_2OH$	Sigma	60-24-2
Triton X 100	$C_{14}H_{22}O(C_2H_4O)_n$ (n = 9-10)	Sigma	9002-93-1
Potassium acetate	CH_3COOK	Sigma	
N Lauryl sarkosyl	$C_{15}H_{30}NNaO_2$	Sigma	137-16-6
Aceticacid	CH_3COOH	Sigma	64-19-7
Guanidium chloride (Aminomethanamidine hydrochloride)	CH_5N_3, HCl	Sigma	
Ethanol (Ethylalcool)	CH_3CH_2OH	Sigma	64-17-5
Silicondioxide	$Si\text{-}O_2$	Sigma	7631-86-9
Proteinase K		Sigma	39450-01-6

EQUIPMENT

Swing-out rotor
Biological hood
Miracloth filter 100 µm; Dominique Dutscher ref 39 3007
Cell Strainer 40 µm; Dominique Dutscher ref 07 4007

Annexe-2 : Liste des clones de Pinots ENTA-INRA® agréés étudiés en S-SAP

Pinot noir n°388
Pinot noir n°111
Pinot noir n°112
Pinot noir n°113
Pinot noir n°114
Pinot noir n°115
Pinot noir n°162
Pinot noir n°163
Pinot noir n°164
Pinot noir n°165
Pinot noir n°236
Pinot noir n°292
Pinot noir n°372
Pinot noir n°373
Pinot noir n°374
Pinot noir n°375
Pinot noir n°386
Pinot noir n°389
Pinot noir n°459
Pinot noir n°461
Pinot noir n°462
Pinot noir n°521
Pinot noir n°528
Pinot noir n°583
Pinot noir n°617
Pinot noir n°665
Pinot noir n°666
Pinot noir n°667
Pinot noir n°668
Pinot noir n°743
Pinot noir n°777
Pinot noir n°778
Pinot noir n°779
Pinot noir n°780
Pinot noir n°792
Pinot noir n°828
Pinot noir n°829
Pinot noir n°870
Pinot noir n°872
Pinot noir n°943
Pinot gris n°457
Pinot gris n°52
Pinot gris n°53
Pinot blanc n°54
Pinot blanc n°55
Pinot Meunier n°900
Pinot Meunier n°924
Pinot Meunier n°818
Pinot Meunier n°819
Pinot Meunier n°916
Pinot Meunier n°925
Pinot Meunier n°865
Pinot Meunier n°458
Pinot Meunier n°901
Pinot Meunier n°817
Pinot Meunier n°791
Pinot Meunier n°864
Pinot Meunier n°977
Pinot Meunier n°983
Pinot Meunier n°978

Annexe-3 : Nombre de copies de chaque élément transposable chez les clones de Pinots étudiés en 454.

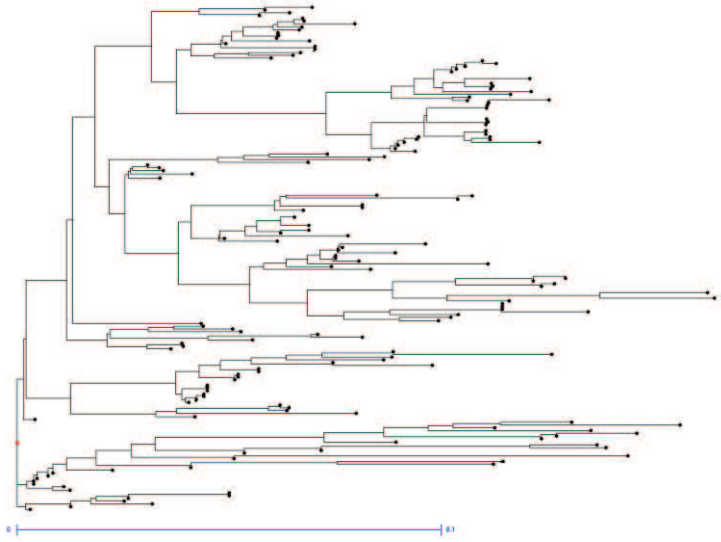
(Tableau de gauche en orange) : Polymorphisme d'insertion entre les clones séquencés en 454 et le PN115 ; (Tableau de droite en bleu) nombre d'insertions contenues dans le génome des clones séquencés en 454. En jaune les éléments étudiés en détail au cours de cette thèse.

Mobile elements ranked in increasing order	Polymorphism copies		Mobile elements ranked in increasing order	Copies of mobile elements from PN386	Copies of mobile elements from PN583	Copies of mobile elements from PN777	mean
VLINE3	7		VLINE1	1796	1476	1723	1665
Gypsy7	6		VLINE3	1661	1246	1647	1518
Gypsy22	6		Copia10	1249	1236	1335	1273
Gypsy17	5		VLINE2	1256	1125	1288	1223
Gret1	5		VLINE6	1548	997	977	1174
VHARB-N1	5		VLINE4	1180	917	1229	1109
Gypsy6	5		MUDRAVI1	933	903	920	919
Copia10	4		Gypsy7	859	894	936	896
Gypsy12	4		Caulimoviridae	1083	889	1223	1065
Gypsy2	4		VHARB-N3	1066	859	1388	1104
MUDRAVI2	4		VLINE5	857	824	1359	1013
VIHAT1	4		Gypsy13	779	809	979	856
Gypsy3	4		VIHAT1	785	699	931	805
Gypsy19	3		Gypsy17	661	670	782	704
Copia-31	3		Gypsy12	1089	668	732	830
GYVIT1	3		MUDRAVI2	706	659	756	707
Gypsy9	3		MUDRAVI2	676	657	813	715
MUDRAVI1	3		Gypsy14	581	586	603	590
VLINE2	3		Gypsy22	615	547	609	590
Copia26	3		GYVIT1	594	531	630	585
VLINE5	3		EnSpm-5	597	509	483	530
VHARB4	3		Gypsy19	564	490	500	518
Gypsy14	2		VIHAT2	467	472	1068	669
Cauliv-1	2		MuDR-21	455	449	469	458
Gypsy4	2		EnSpm-4	484	431	483	466
Gypsy13	2		Gypsy4	422	403	683	503
VLINE1	2		Harbinger-3	454	403	489	449
Copia22	2		MuDR-6	367	381	474	407
EnSpm-5	2		hAT-10	456	378	397	410
VHARB-N3	2		MuDR-13	526	370	526	474
Harbinger-1	2		Gypsy20	395	350	376	374
Copia1	2		MuDR-8	361	349	407	373
VIHAT3	2		Gypsy9	379	340	332	350
Copia-35	2		VHARB4	344	326	349	340

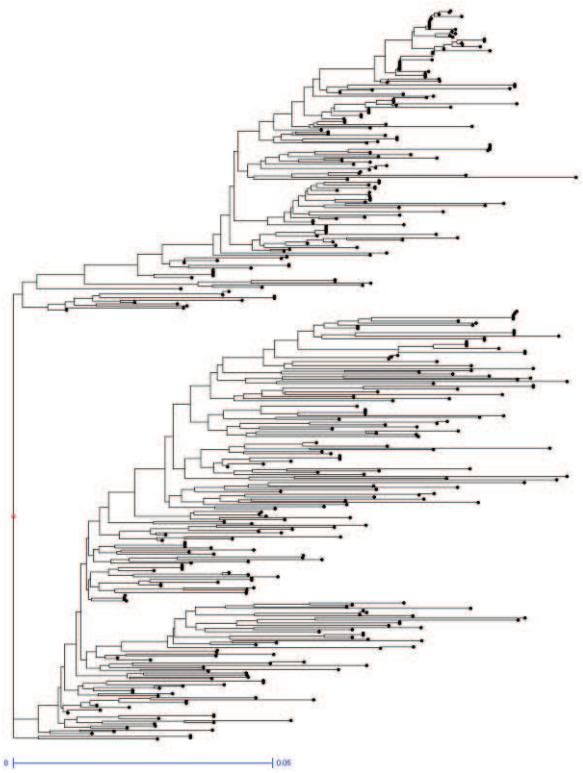
Copia15	2		ENSPM2	421	321	541	427
Copia8	2		Gypsy11	427	314	317	353
Copia23	1		Gypsy18	291	277	276	281
Copia11	1		MuDR-9	301	266	301	289
Gypsy11	1		MuDR-9	324	249	279	284
Copia17	1		Gypsy3	260	230	311	267
Harbinger-3	1		EnSpm-13	239	228	270	246
VLINE6	1		MuDR-12	266	222	544	344
VLINE4	1		VIHAT3	209	221	366	265
Copia-32	1		Gypsy6	220	213	406	279
VIHAT2	1		MuDR-18	274	196	209	227
MuDR-18	1		Harbinger-1	200	196	215	204
Copia16	1		Gypsy2	193	189	221	201
Copia24	1		MuDR-4	185	188	166	180
Copia27	1		EnSpm-3	203	185	178	189
hAT-10	1		ENSPM-N3	207	182	219	202
MuDR-3	1		MuDR-3	219	181	186	195
Gypsy18	1		hAT-6	213	173	279	222
Copia25	1		ENSPM1	187	170	183	180
Copia18A	1		EnSpm-6	183	164	177	175
Copia3	1		hAT-7	173	154	179	169
EnSpm-3	1		VHARB-N2	155	146	155	152
Copia19	1		hAT-11N	117	114	183	138
ENSPM-N3	1		Vine-1	119	113	169	134
hAT-6	1		Copia-31	109	99	109	106
Copia1A	1		Gypsy16	118	92	145	118
ENSPM1	1		MuDR-5	87	79	75	80
Copia5	1		Copia-33	78	73	147	100
VLINE3	7		MuDR-11N	81	69	75	75
Gypsy7	6		Gret1	64	66	64	65
Gypsy22	6		Copia9	69	51	82	67
Gypsy17	5		Copia26	41	44	117	67
Gret1	5		Copia23	49	43	49	47
VHARB-N1	5		Copia-29	37	35	82	51
Gypsy6	5		Copia22	38	31	50	40
Copia10	4		Copia17	34	27	29	30
Gypsy12	4		Helitron1	23	26	32	27
Gypsy2	4		Copia3	19	23	62	35
MUDRAVI2	4		Copia11	23	19	17	20
VIHAT1	4		MuDR-7	20	19	21	20
Gypsy3	4		Copia12	27	19	17	21
Gypsy19	3		Copia15	21	18	25	21
Copia-31	3		Copia5	16	17	25	19
GYVIT1	3		EnSpm-8N	17	16	35	23
Gypsy9	3		Tvv1	18	16	35	23

MUDRAVI1	3		Copia18A	11	11	15	13
VLINE2	3		Copia18	8	10	31	16
Copia26	3		Gypsy1	6	9	9	8
VLINE5	3		Copia-34	7	8	22	12
VHARB4	3		Copia-32	7	8	7	7
Gypsy14	2		Copia1	9	8	22	13
Cauliv-1	2		Copia16	7	6	19	11
Gypsy4	2		Copia28	3	5	9	6
Gypsy13	2		Copia29	3	5	8	5
VLINE1	2		Copia-35	6	4	14	8
Copia22	2		Copia7	3	4	9	5
EnSpm-5	2		Copia24	6	3	8	6
VHARB-N3	2		Gypsy8	3	3	9	5
Harbinger-1	2		Copia2	2	3	5	3
Copia1	2		Copia25	3	3	9	5
VIHAT3	2		Gypsy15	2	3	5	3
Copia-35	2		Copia27	3	2	9	5
Copia15	2		Copia19	1	2	3	2
Copia8	2		Copia-30	2	2	4	3
Copia23	1		Copia20	2	2	6	3
Copia11	1		Gypsy5	1	1	4	2
Gypsy11	1		Copia6	1	1	4	2
Copia17	1		Copia1A	0	1	2	1
Harbinger-3	1		Copia13	1	1	2	1
VLINE6	1		Copia8	1	1	1	1
VLINE4	1		Gypsy10	1	1	3	1
Copia-32	1		Copia21	1	0	1	1
VIHAT2	1		Copia4	0	1	1	1

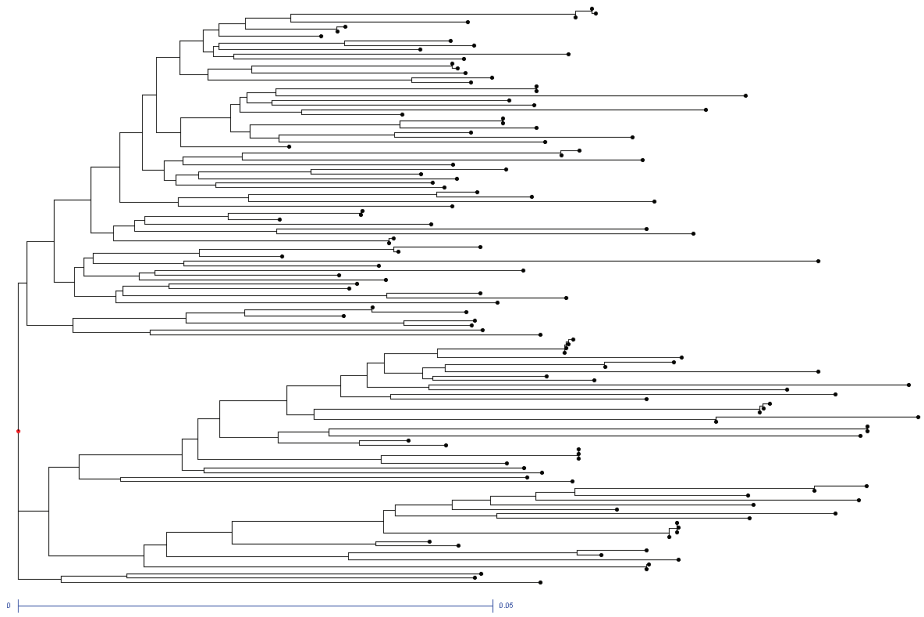
Annexe-4 : Arbre d'homologie des différentes formes d'insertion de trois éléments transposables.



Supplementary Figure 4. Caul-1 phylogenetic tree.



Supplementary Figure 5. Copia-10 phylogenetic tree.



Supplementary Figure 6. Gypsy-19 phylogenetic tree.

Annexe-5 : Liste cépages étudiées en S-SAP. Chaque couleur correspond à une espèce de Vitis.

Details of the S-SAP protocol

S-SAP was used to validate the polymorphism of mobile elements. To set up this protocol, we referred to previous studies (Knox *et al.* 2009, Labra *et al.* 2004, Wegscheider *et al.* 2009). DNA extracted from four Pinot noir clones (150 ng) was restricted with MseI (InvitroGen). Digestion was performed for 10 h at 65 °C. DNA ligation of *MseI* adapter was prepared by adding (5X) Ligase buffer, 400 U of T4 DNA ligase, 25 pMol of *MseI* adapter and incubated for 4 h at room temperature. T4 Ligase was inactivated by heating at 65 °C for 10 min and samples were stored at 4 °C. Ligated DNA was diluted 1;10 and pre-amplified with 10 mM of *MseI* primer (20 ng DNA, 1X PCR buffer, 3 mM MgCl2, 2.5 mMdNTPs and 1 U Taq DNA polymerase) with the following program: one denaturation step at 94°C for 3 min, following by 30 amplification cycles (94 °C for 45 sec, 56 °C for 45 sec, 72 °C for 60 sec) and a final elongation step at 72 °C for 3 min. Selective amplification was performed with 5 µL of pre-amplified DNA, 1X PCR buffer, 3 mM MgCl2, 2.5 mMdNTPs, 10 pM primer of retrotransposon fluorescent markers (FAM or HEX), 10 mM of specific primer of *MseI* at 3 selective bases (1/16 of the genome) (5' GAT GAG TCC TGA GTA ACG T) and 1U Taq DNA polymerase. Using PCR program: a denaturation step at 94 °C for 3 min, followed by 30 amplification cycles using a touchdown from 65 to56 °C (94 °C for 45 sec, 65 °Cf or 45 sec with a touchdown of -0.7°C per cycle, 72 °C for 60 sec) followed by 25 amplification cycles (94 °C for 45 sec, 56 °C for 45 sec, 72 °C for 60 sec) and a final elongation step at 72 °C for 3 min. A 1:10 dilution of the fluorescently labeled amplified DNA fragments was run on the Applied Biosystem 3730 xl with the ladder size standard set at 524 bases. The S-SAP profile was analyzed using GeneMapper® (Applied biosystems). We considered there was a peak if the threshold exceeded 150 lux.

Annexe-6 : Liste cépages étudiées en S-SAP. Chaque couleur correspond à une espèce de Vitis.

Nom d'introduction	Espèce	PaysOrigine	Nom d'introduction	Espèce	PaysOrigine
Tsolikouri	vinifera	Géorgie	Vitis riparia bois rouge	riparia	Amériques
Voskeat	vinifera	Arménie	Riparia Gloire de Montpellier	riparia	Amériques
Kapistroni_tétri_hermaphrodite_(Coll._Kichinev)	vinifera	Géorgie	Vitis riparia à lobes acuminés	riparia	Amériques
Espadeiro_tinto	vinifera	Portugal	Vitis riparia à bourgeons bronzés n°2	riparia	Amériques
Plant_du_Maroc_E_(Collection_Meknès)	vinifera	Algérie	Vitis riparia à bourgeons dorés	riparia	Amériques
César	vinifera	France	Vitis riparia Baron Périer	riparia	Amériques
Orlovi_nokti	vinifera	Roumanie	Vitis riparia Scribner	riparia	Amériques
Araklinos	vinifera	Grèce	Vitis riparia Meissner n°8	riparia	Amériques
Lameiro	vinifera	Portugal	Vitis riparia à pousse vineuse	riparia	Amériques
Medouar	vinifera	Israel	Vitis riparia Pulliat	riparia	Amériques
Chirai_obak	vinifera	Tadjikistan	Vitis riparia géant de Las Sorres	riparia	Amériques
Tsitsa_Kaprei	vinifera	Roumanie	Vitis riparia Meissner n°1	riparia	Amériques
Lambrusque_Akchour_S21_Ind_4	sylvestris	Algérie	Vitis berlandieri Boutin B	berlandieri	Amériques
Lambrusque_Grésigne_1	sylvestris	France	Vitis berlandieri Chiendent	berlandieri	Amériques
Lambrusque_I	sylvestris	France	Vitis berlandieri n°1 Salomon	berlandieri	Amériques
Lambrusque_Sejnene_1	sylvestris	Tunisie	Vitis berlandieri n°5 Salomon	berlandieri	Amériques
Silvestris_de_Draa_Ben_Khedda	sylvestris	France	Vitis berlandieri Resseguier n°1	berlandieri	Amériques
Vigne_de_Pausanias	sylvestris	France	Vitis berlandieri Las Sorres n°9	berlandieri	Amériques
Vigne_sauvage	sylvestris	Slovaquie	Vitis berlandieri Mazade	berlandieri	Amériques
Vitis_silvestris_3-23	sylvestris	France	Vitis berlandieri n°10	berlandieri	Amériques
Vitis_silvestris_Afghanistan	sylvestris	France	Vitis cinerea Arnold	cinerea	Amériques
Vitis_silvestris_Guemuelder_104-64_mâle	sylvestris	France	Vitis cinerea Davin	cinerea	Amériques
Vitis silvestris Ketsch 2-39 mâle	sylvestris	Allemagne	Vitis cinerea Illinois	cinerea	Amériques
Lambrusque Ul any nad Zitavou A77	sylvestris	Slovaquie	Vitis cinerea Canescens	cinerea	Amériques

Résumé

L'exploitation de la variation clonale est une des voies d'amélioration utilisée chez un grand nombre de plantes d'intérêts agronomiques telles que la pomme de terre, le café et la vigne. En effet, après plusieurs cycles de reproduction végétative, des caractéristiques agronomiques stables apparaissent donnant naissance à une diversité phénotypique remarquable, appelée « diversité clonale ». Chez la vigne, cette diversité clonale est d'une importance majeure pour les viticulteurs puisqu'elle permet une amélioration variétale sans changer d'identité de cépage en conformité avec la réglementation fixée par Appellations d'Origine Protégée.

L'hypothèse la plus parcimonieuse expliquant cette diversité phénotypique clonale est l'accumulation de mutations somatiques au cours des cycles de reproduction végétative. L'objectif de cette thèse a été de dresser un panorama le plus exhaustif possible des différents polymorphismes moléculaires entre les génomes de plusieurs clones. Dans un premier temps trois clones de Pinot ont été séquencés par la technique 454 GS-FLX puis dans un second temps 11 clones de quatre cépages ont été séquencés la technique Illumina HiSeq 2000. Afin d'analyser la grande quantité de données obtenues, nous avons construit un pipeline d'analyse (Bacchus pipeline) permettant d'identifier tous les types de polymorphismes moléculaires entre les différents génomes.

Nos résultats permettent, pour la première fois un inventaire exhaustif des polymorphismes moléculaires dans un contexte multiplication végétatif. L'ensemble des mutations polymorphes entre deux clones a pu être identifié, SNPs, indels (2,5 SNPs et 11,5 indels par Mb en moyenne) ainsi que des variations d'ordre structural (larges insertions ou délétions) représentant la classe la plus fréquente (129 évènements par Mb entre deux clones en moyenne). Afin d'évaluer le polymorphisme d'insertion généré par ces éléments nous en avons étudié quatre par une approche S-SAP sur plusieurs niveaux de diversité (inter-espèces, inter-cépages, inter-clones et entre plusieurs tissus d'un même individu). L'analyse phylogénétique au niveau des espèces est conforme à celle réalisée avec d'autres types de marqueurs moléculaires (SSR, SNP). Cependant, une forte instabilité de ces insertions a été confirmée entre les clones et entre les tissus d'un même d'individu.

L'identification des clones par une méthodologie moléculaire serait d'une grande importance pour la filère. Pour cet objectif, nos résultats indiquent que les mutations de types SNP et petits indels qui sont certes moins fréquentes que les variations structurales mais qui sont plus stables semblent plus pertinentes pour la mise en place d'une méthodologie d'identification des clones.

Abstract

Clonal variation is considered as an effective contribution to breeding programs of vegetatively propagated species with major agronomical interest such as banana, potato, coffee and grape. Indeed, after several propagation cycles, stable and heritable phenotypic variations appear giving rise to a phenotypic variation termed "clonal diversity". This clonal diversity is very important for wine-growers because it allows preserving cultivars identity in the strict respect of Appellation (A.O.P) wines specifications

The most parsimonious hypothesis explaining clonal phenotypic diversity is the accumulation of somatic mutations. The objective of my thesis was to provide a broad description of molecular polymorphisms in the context of vegetative propagation. Three clones were first sequenced by 454 GS-FLX technology and eleven clones were then sequenced with Illumina Hiseq2000 technique. To analyse the high quantity of data obtained, we built a pipeline (Bacchus pipeline) allowing the identification of all existing molecular polymorphisms between different genomes.

All polymorphism types were observed: indels and SNPs which have a low polymorphism frequency (2.5 SNPs and 11.5 indels per Mb between two clones in average) and structural variations (large insertions or deletions) which have a high polymorphism frequency (129 per Mb between two clones in average) but are unstable. To evaluate stability and polymorphism level of these transposable elements, we have studied 4 elements using S-SAP method at different diversity levels (inter-species, inter-cultivars, inter-clones and between organs/tissues of a single individual). Our interspecific phylogenetic analysis is similar to other phylogenies performed with SSR or SNPs markers. However, we confirm the high instability of these elements between clones and between tissues in single individuals.

Clone identification through molecular methods would be of high significance for the wine industry. SNP or small indels mutations are less frequent but more stable than structural variation and could be used for accurate clone identification.