

Génomique de la vigne et applications

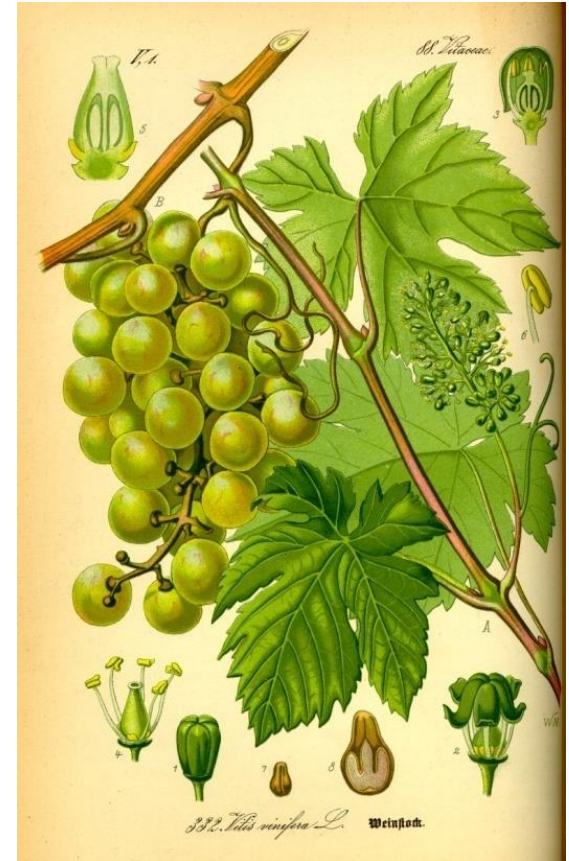
A-F Adam-Blondon



Unité de Recherches
Génomique-Info

Master 1 Vigne et Terroirs, Dijon, 09 mai 2016

Sandrine Balzergue, INRA, Angers
Loïc LeCunff, IFV, Montpellier



PLAN

1- GENERALITES SUR LE GENOME DE LA VIGNE

2- CARTE GENETIQUE

3- SEQUENCAGE DU GENOME DE LA VIGNE

4- **GENOMIQUE FONCTIONNELLE**

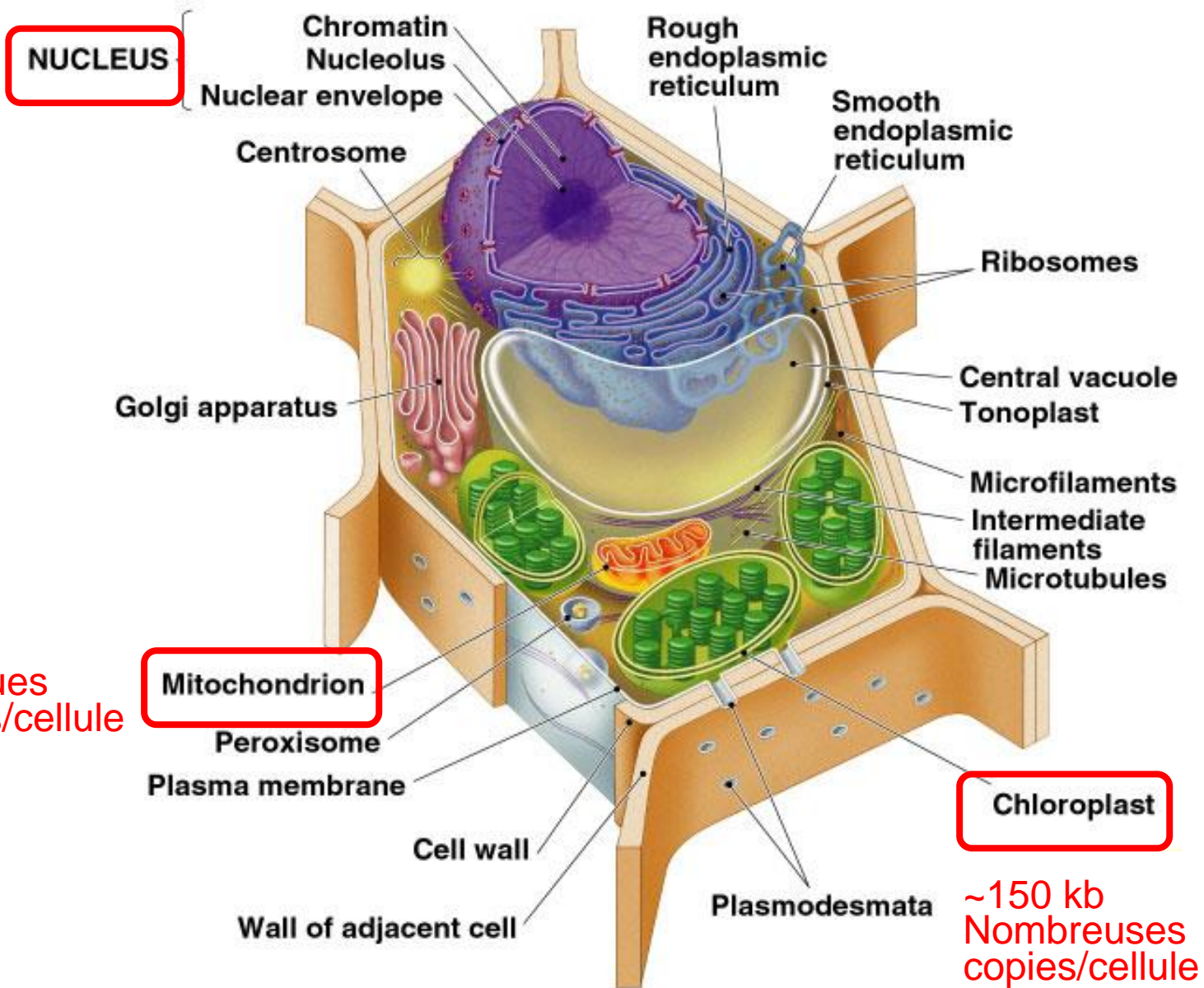
5- CHAMPS D'APPLICATIONS

1

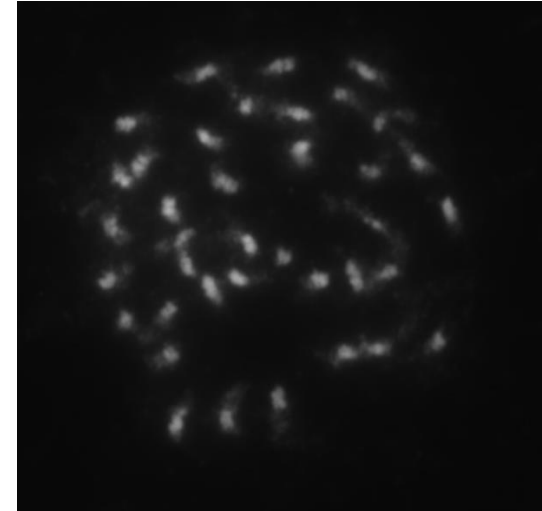
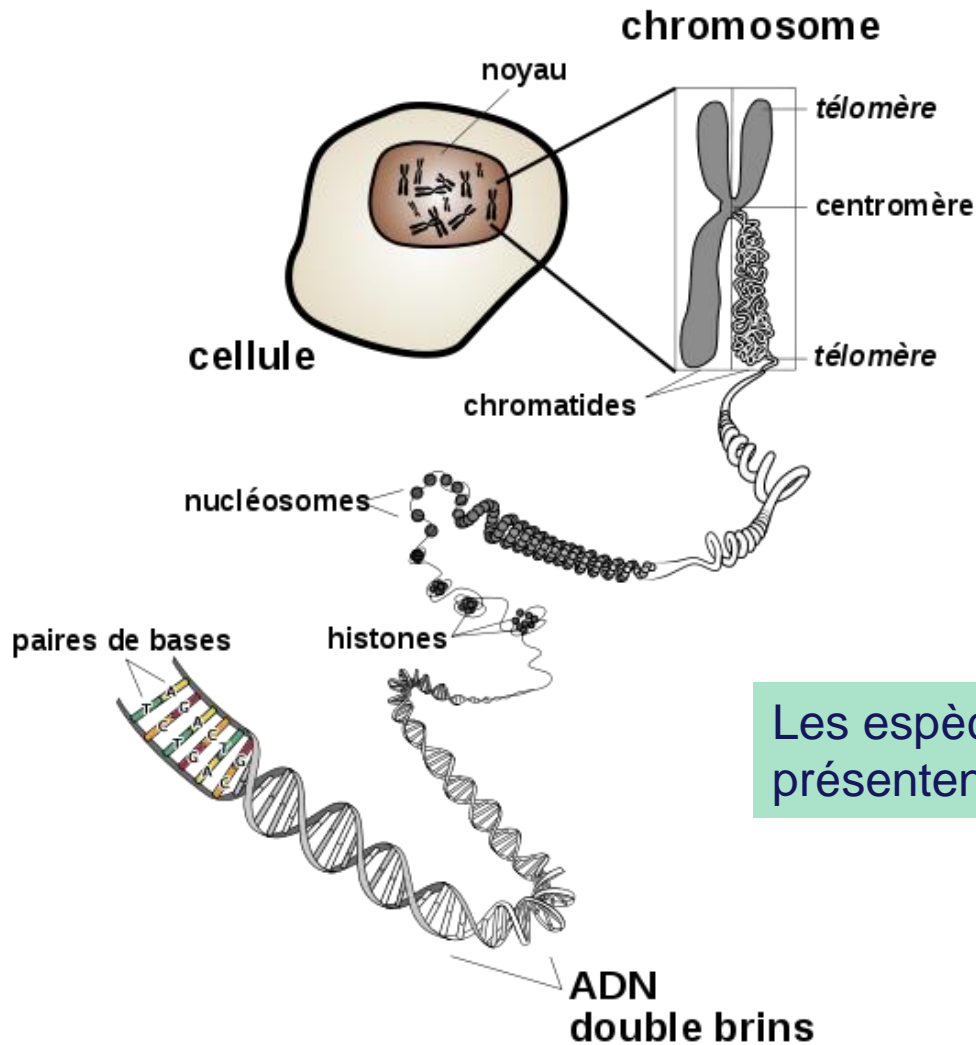
Généralités sur les génomes et le génomme de la vigne

La cellule de vigne et l'extraction d'ADN

~500 Mb
2n=38

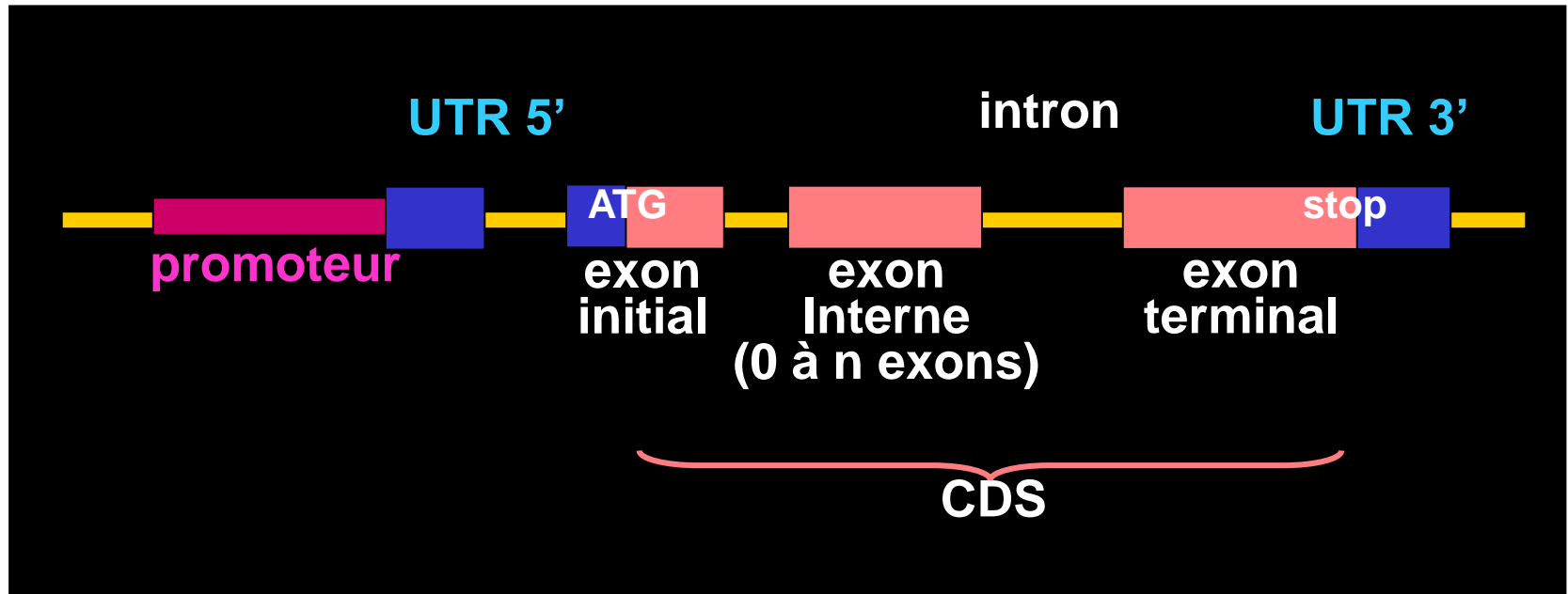
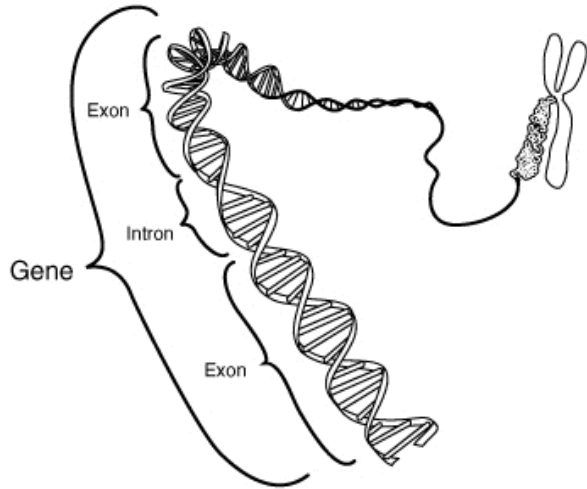


Les espèces du sous-genre *Euvitis* ont 19 paires de chromosomes

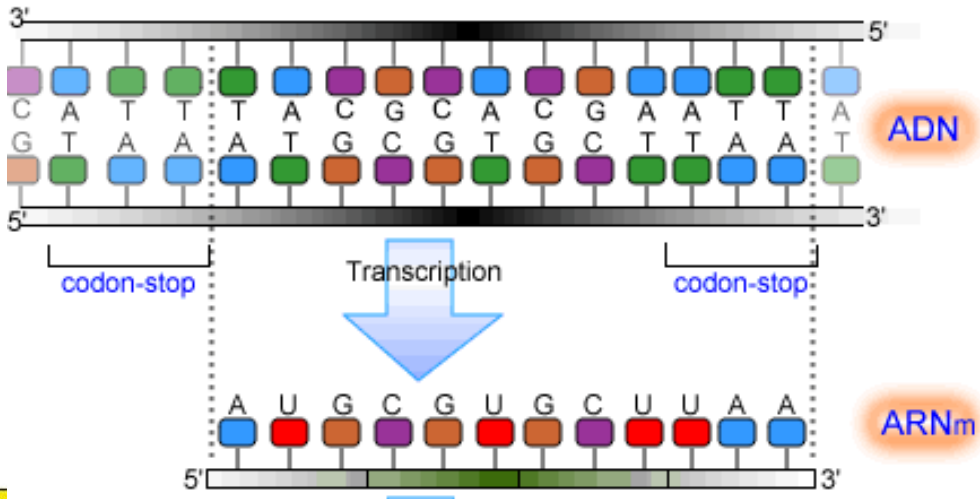


Les espèces du sous genre *Muscadinia* présentent 20 paires de chromosomes

Structure des gènes des Eucaryotes



De la transcription à la traduction des gènes

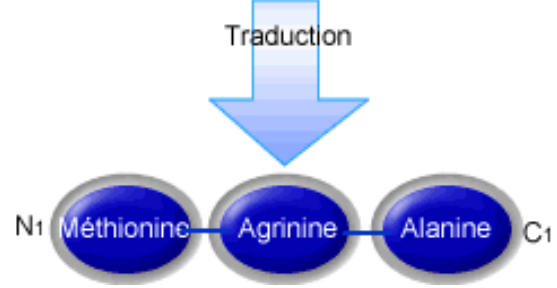


couples de nucléotides possibles :

- T A -
- C G -

couples de nucléotides possibles :

- U A -
- C G -



20 acides aminés possibles

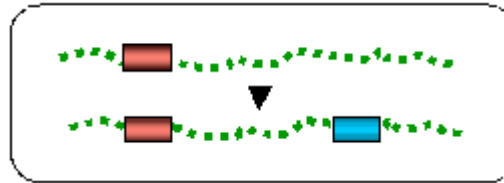
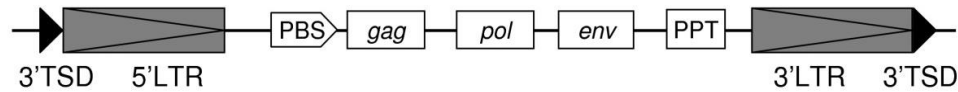
		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G

Les génomes contiennent également des séquences répétées: les éléments transposables en majorité



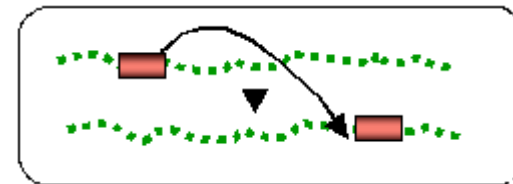
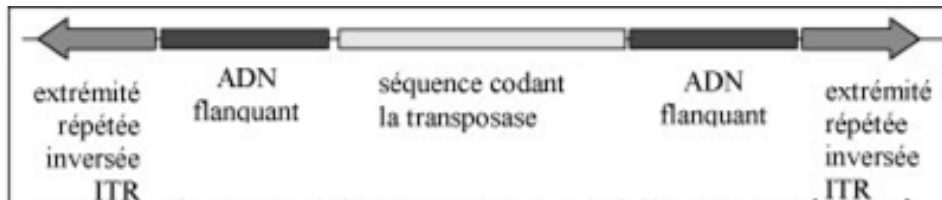
Classe I: rétro-transposons

Intermédiaire ARN/cDNA : « copier-coller »



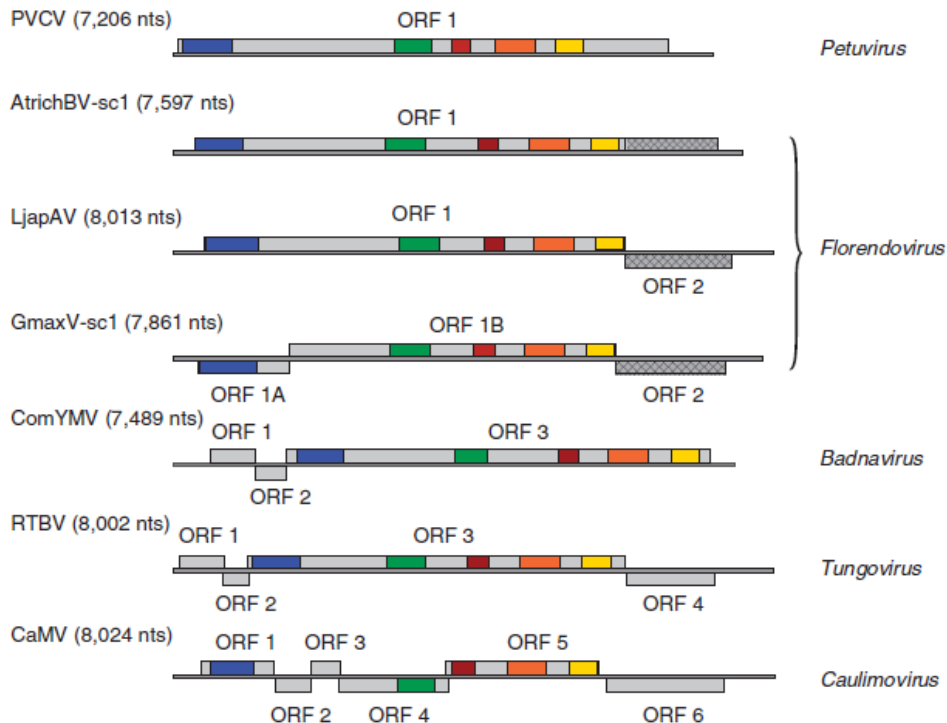
Classe II: transposons

« Couper-Coller »



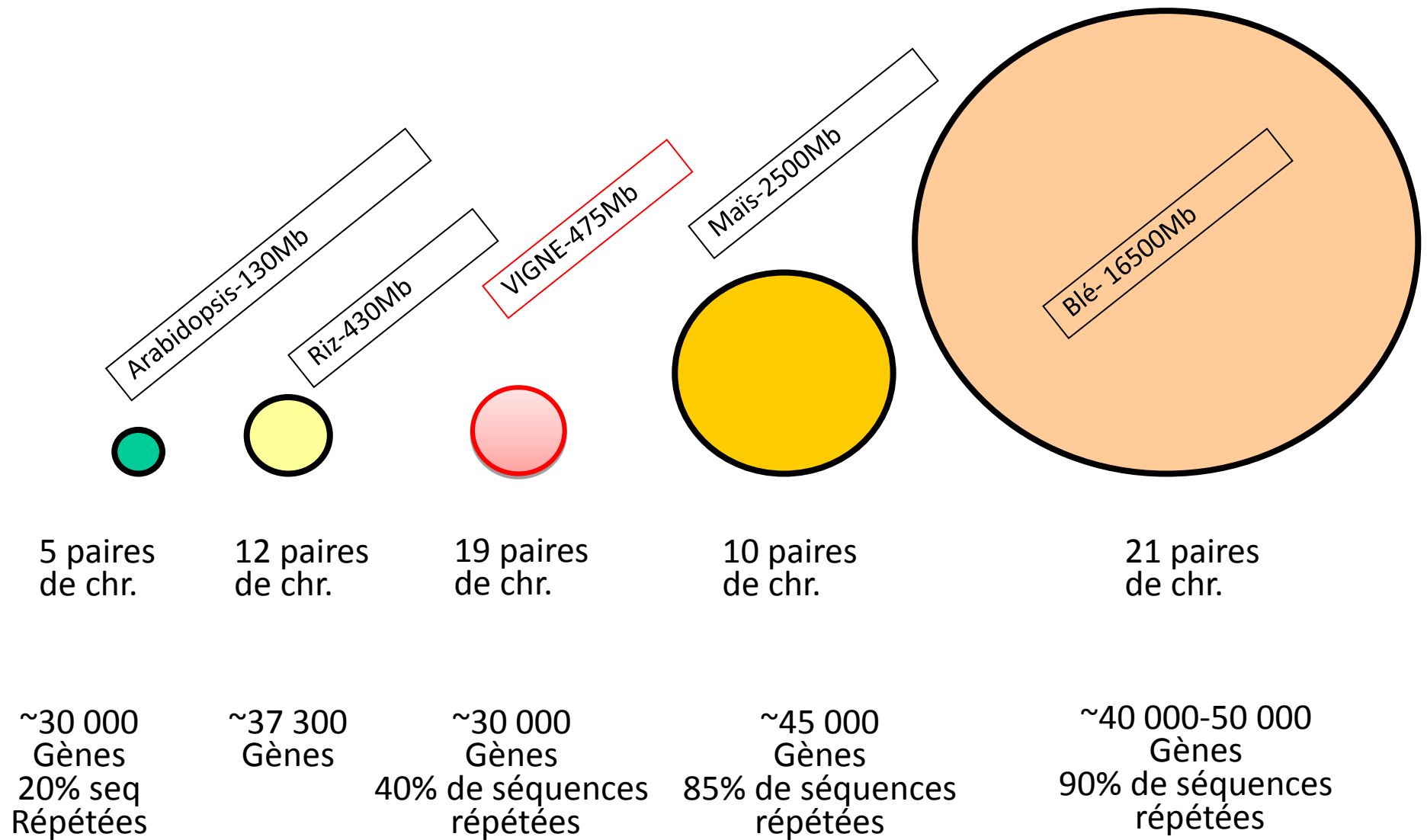
Les génomes contiennent également des séquences répétées: les virus endogènes

- Des séquences de virus sont parfois intégrées dans les génomes des eukaryotes et évoluent comme des séquences répétées (Endogeneous Virus Elements = EVE; Freschotte et Gilbert -2012- Nat Rev Genet, 13: 283-296)
- Ces séquences sont dérivés de la famille des Caulimovirus chez les plantes à fleurs et représentent environ 0.65% du génome de la vigne: Florendovirus (Geering et al -2014- Nature Comm, DOI:10.1038/ncomms6269)

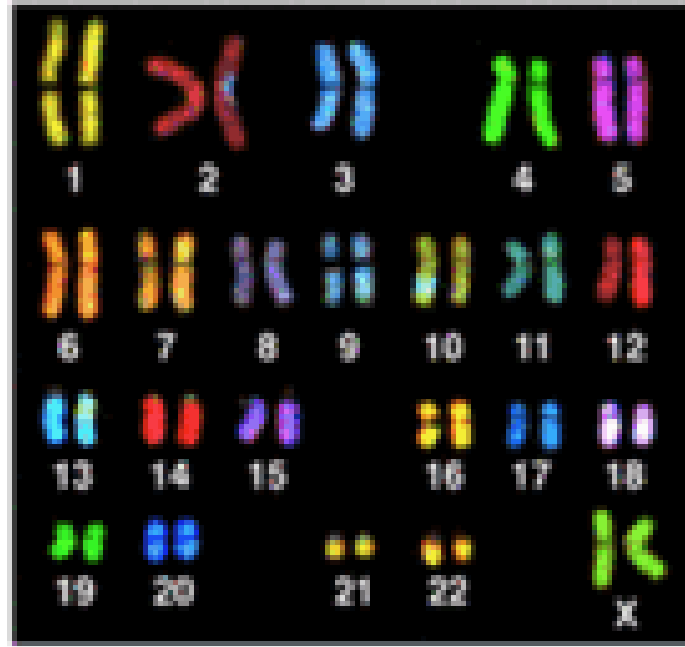


9% des Florendovirus de la vigne sont situés dans des introns => rôle dans la régulation de l'expression des gènes?

La taille des génomes de plante varie considérablement en fonction de leur quantité d'élément transposable et ...



...de leur niveau de ploïdie



Le nombre de chromosomes portant la même information génétique dans une cellule varie en fonction de l'espèce étudiée. Chez la vigne, l'humain, le chien, le chat etc. il est de 2: ces espèces sont dites diploïdes.

Lorsqu'il est supérieur à 2 comme chez la canne à sucre, l'igname ou le blé, ces espèces sont dites polyploïdes

La vigne a un génome de petite taille et diploïde => favorable à des approches génomiques

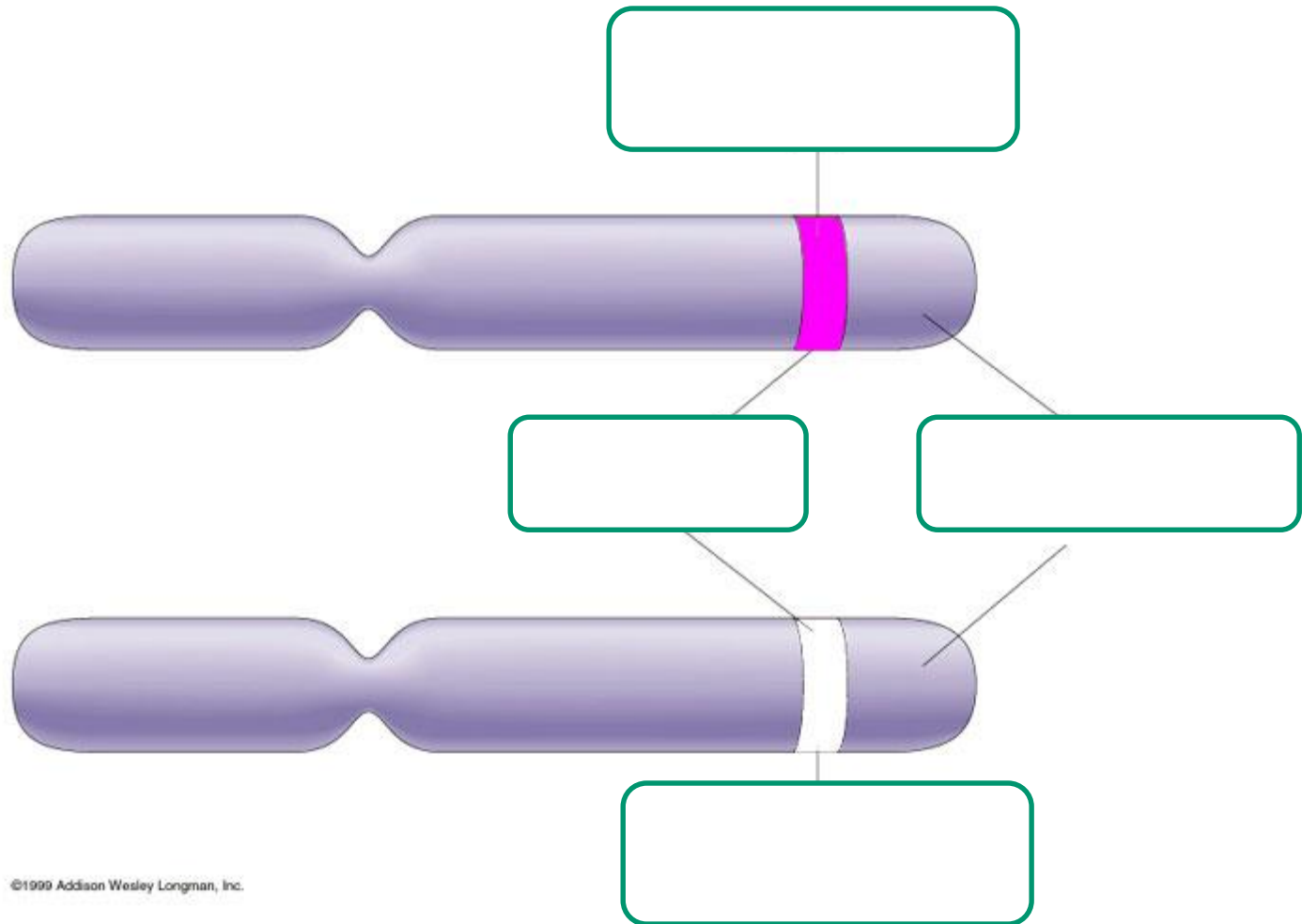
La génomique

- Etude de la structure, du polymorphisme et du fonctionnement des génomes avec des approches globales
- Notion de haut débit du point de vue technologique
- Importance de l'informatique et de la bio-informatique
- Importance des statistiques

2

Les cartes génétiques

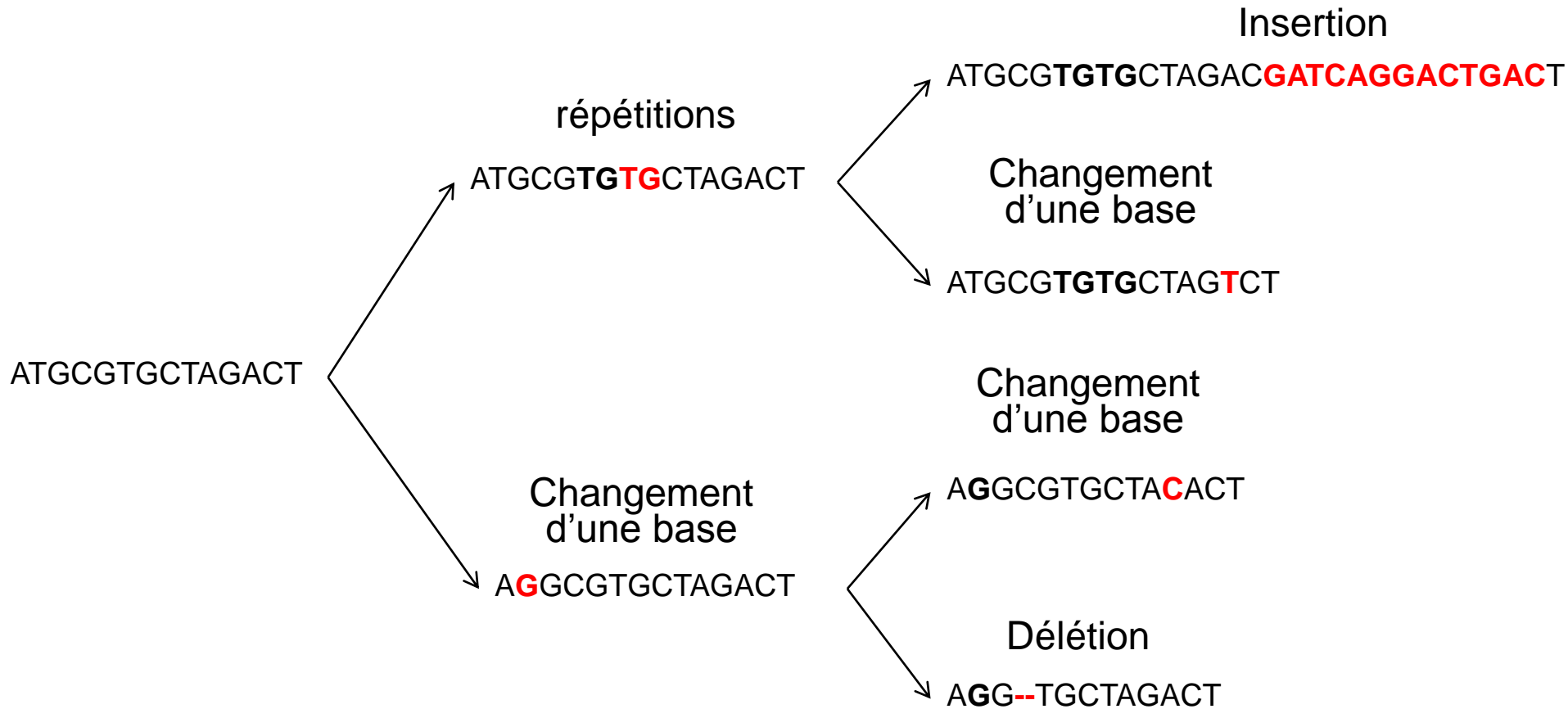
La vigne a un génome très hétérozygote



©1999 Addison Wesley Longman, Inc.

Remplacez les termes : Locus, Allèle, homozygote, hétérozygote

Les différents types de polymorphisme de séquence observables à un locus donné

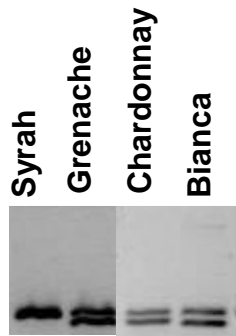


Le polymorphisme de séquence et les marqueurs génétiques

A

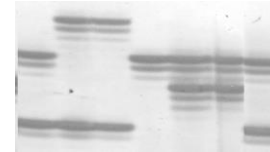
ATACGATTTTATAAT **AGAGAGAGAGAGAGAGAGAGAGAGAGAGAG** ATGTGGGATGTG

B



Marqueur SC8_0108_086

Mauzac
Clairette
Mourvèdre
Tannat
Colombard
Riesling
Ugni blanc



Marqueur VVMD5

Quel type de séquence est représenté en encadré A?

Qu'est ce qui est représenté en encadré B?

Y-a-t-il une variété homozygote/hétérozygote?

Utilisation de ce type de marqueurs?

Autres types de marqueurs?

Les marqueurs moléculaires

- Dominant ou co-dominants
- Locus spécifiques ou non

•**Marqueurs codominants spécifiques:** Microsatellites, SNP

•**Marqueurs dominants non locus spécifiques:** AFLP, M-SAP, S-SAP

Microsatellites : variation du nombre de répétitions du motif

individu 1: GG**TGGTGG**TTC => (**TGG**)₂

individu 2 : GG**TGG**TTC => (**TGG**)₁

individu 3 : GG**TGGTGGTGG**TTC => (**TGG**)₃

Polymorphisme Nucléotidique Simple (SNP)

individu 1: GGAT**C**TTC

individu 2 : GGAT**A**TTC

Les marqueurs microsatellites vs SNP

Microsatellites

Avantages :

- Séquences très polymorphes
- Co-dominants
- Couverture intéressante du génome
- Manipulables par PCR
- Automatisables
- Haut débit

Inconvénients:

- Développement spécifique obligatoire
- Séquences non codantes/marqueurs neutres

SNP

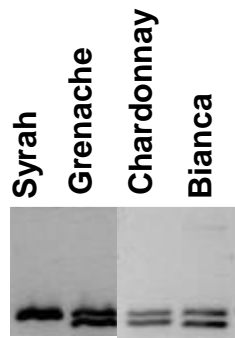
Avantages :

- Extrêmement fréquents
- Très bien répartis sur le génome
- Co-dominants
- Manipulables par PCR
- Automatisables
- Très haut débit

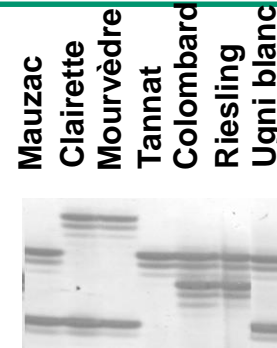
Inconvénients:

- Développement spécifique obligatoire
- Limité à 4 (2) allèles par locus

Polymorphisme dans les ressources génétiques vigne



Marqueur SC8_0108_086



Marqueur VVMD5

Espèces du genre *Vitis*:
V. riparia, *V. rupestris*...
Muscadinia rotundifolia

Portabilité, allèles
spécifiques des espèces

SNP: 2% de marqueurs
transférables => travail avec
des « bouquets » de
marqueurs très proches

Vitis vinifera cultivés
Vitis vinifera sylvestris
(lambrusques)

Hétérozygotie et variation
allélique importantes

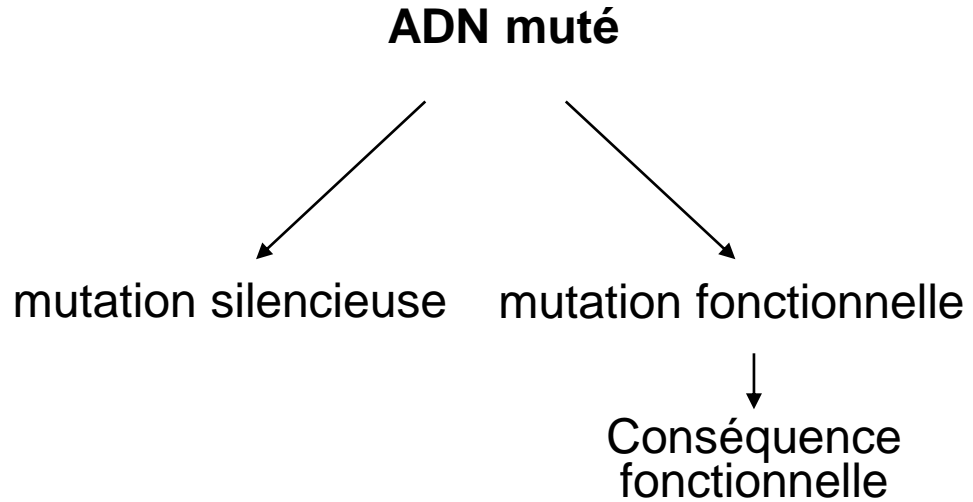
SNP: probablement
plusieurs dizaines

Collections de clones
des principales variétés
de *Vitis vinifera*

Variation allélique très faible

Pas le mécanisme principal
de variation clonale

Impact d'une mutation / polymorphisme



Substitution d'un nucléotide dans le gène *vvGAI* => changement d'une Leucine en Histidine

Grapevine

VvGAI1

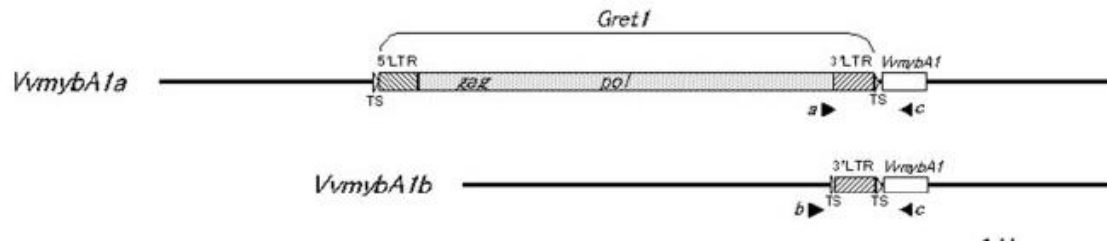
Vvgai1

AGMDEL I **L**AVLGYNVKASDMAEVAQKLEQLEEVI VNAQEDGLSHLASETVHYNPSDLSNWLGSMLS
AGMDEL **H**AVLGYNVKASDMAEVAQKLEQLEEVI VNAQEDGLSHLASETVHYNPSDLSNWLGSMLS

Boss et Thomas, Nature, 2002

Impact d'une mutation / polymorphisme

Gène *Myb* impliqué dans la variation « on/off » de la couleur de la baie

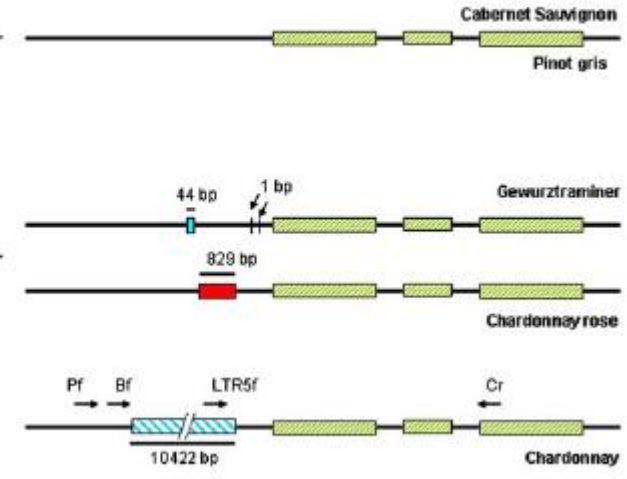


Kobayashi et al , Science, 2004

Walker et al, Plant J, 2007

Dans une collection de 141 variétés de vignes la présence de couleur est toujours associée à l'absence d'un rétrotransposon dans le promoteur du gène *VvMybA1*

<i>Gret1</i> absent	<i>Gret1</i> présent	total	
65	0	65	Noir
14	0	14	Gris
17	0	17	Rouge
18	0	18	Rose
3	81	84	Blanc



This et al, TAG, 2007

Recherche d'associations entre le polymorphisme d'ADN et des variations de caractères d'intérêt

Objectifs:




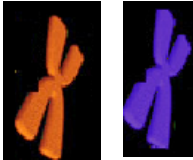
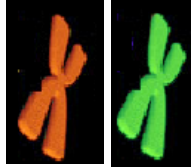

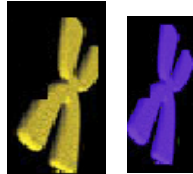
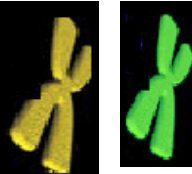
- Disposer de marqueurs prédictifs du phénotype
- Identification des polymorphismes responsables de la variation du phénotype => compréhension des mécanismes moléculaires et physiologiques

Approches *de novo*:

- Cartographie génétique
- Génétique d'association

Utilisation des marqueurs moléculaires pour l'établissement de cartes génétiques



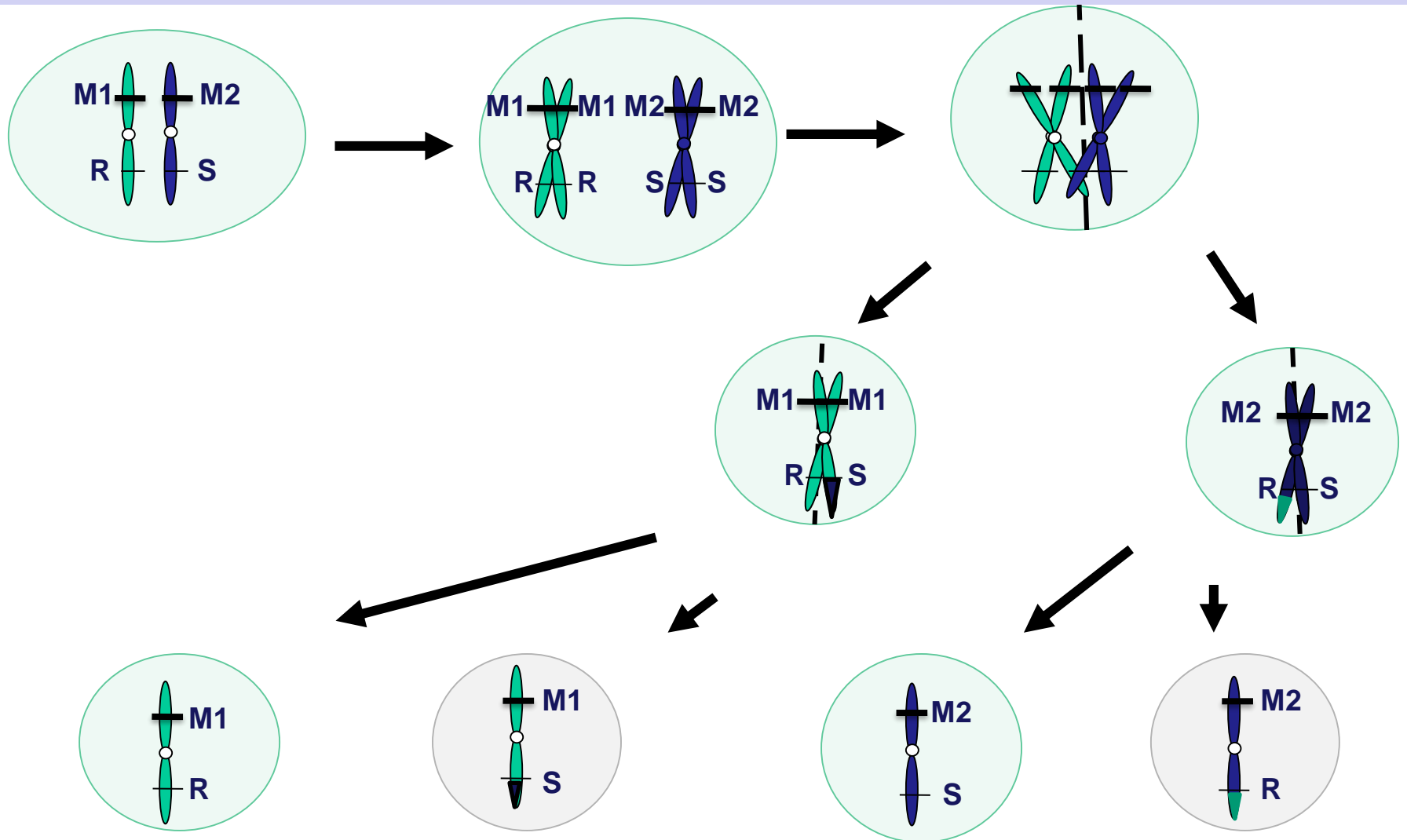
		Gamètes mâles	
			
Gamètes femelles			
			

Chez les espèces hétérozygotes on travaille en général avec des familles « pleins-frères »

Ségrégation des allèles chez la mère et chez le père

Que peut-il se passer dans le cas d'autofécondations?

Les cartes génétiques sont basées sur l'observation de la ségrégation de marqueurs/caractères polymorphes dans des descendance



Distance génétique entre deux marqueurs = taux de recombinaison « r »

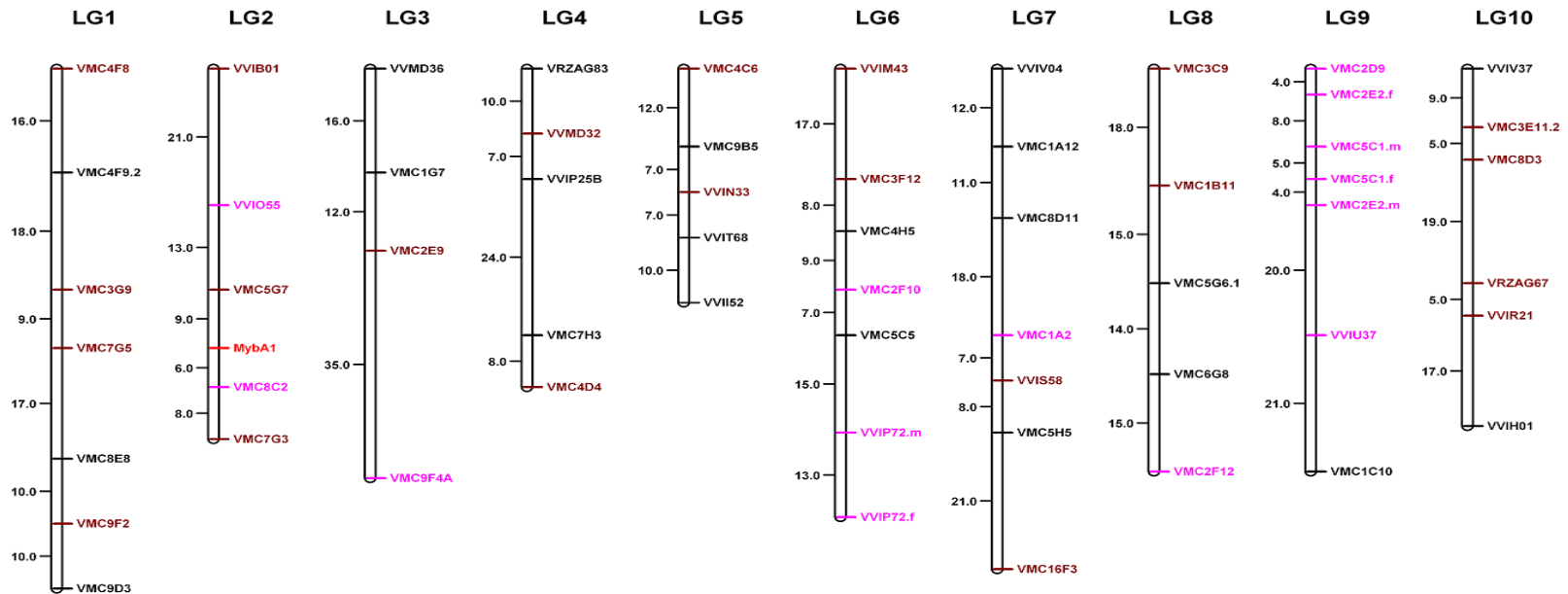
$$r = \frac{\text{Nombre de gamètes recombinants}}{\text{Nombre de gamètes totaux}}$$

$$0 < r < 0,5$$

locus complètement liés: $r = 0$ locus indépendants: $r = 0,5$

Plus les marqueurs physiquement éloignés, plus la probabilité d'occurrence d'une recombinaison entre eux est grande => distance élevée

Carte génétique



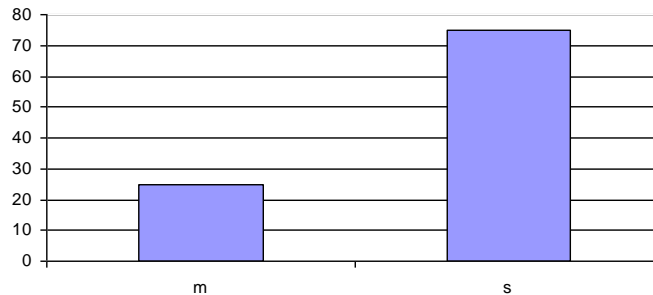
Croisement Syrah X Grenache

191 individus

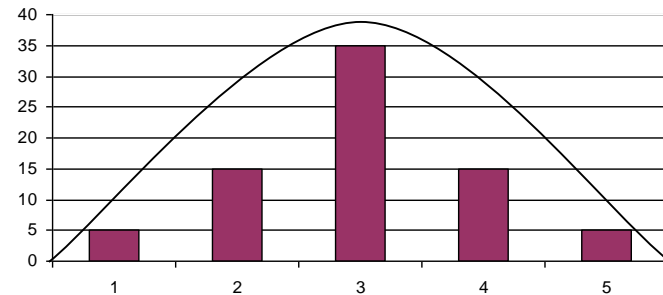
Carte génétique ~1500 cM

Localisation des régions des génomes impliqués dans la variation de caractères agronomiques par cartographie génétique

Caractère qualitatif



Caractère quantitatif



Localisation d'une région du génome impliquée dans la variation d'un caractère (Quantitative Trait Loci ou QTL):

- pour chaque marqueur de la carte comparaison de la moyenne au caractère des individus portant l'allèle A et de ceux portant l'allèle B
- Si les moyennes sont significativement différentes, le locus est dans une région qui contrôle le caractère

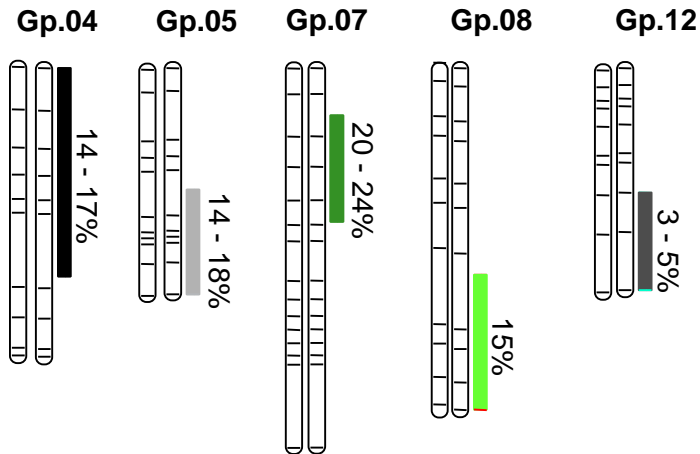
Cartographie génétique: première approche génomique pour faire le lien entre le génotype et le phénotype

génotype

Expression des gènes

Environnement

phénotype

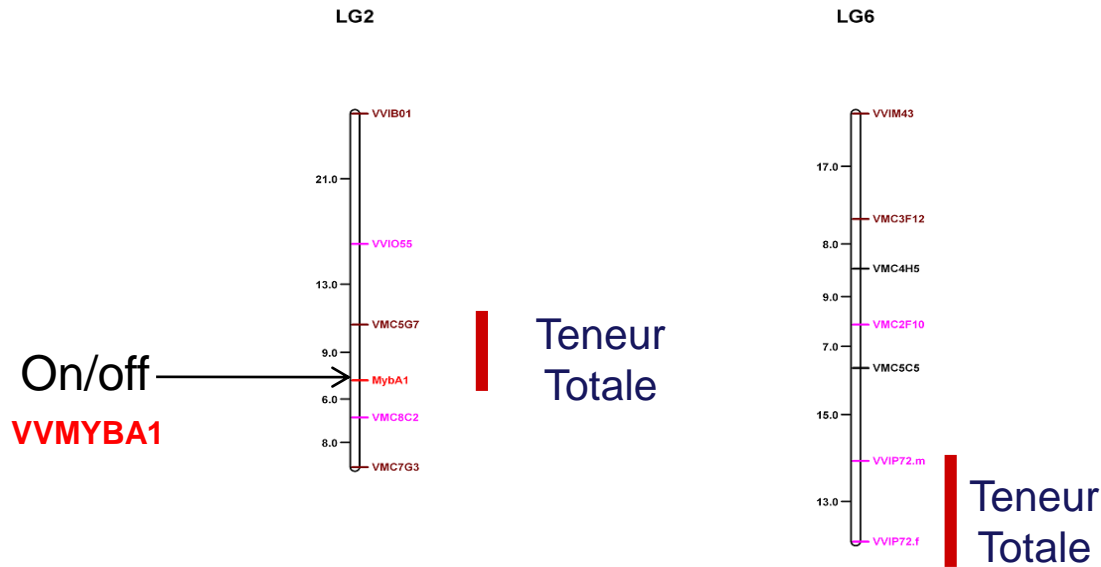


Longueur de la grappe, Compacité, Poids baie
Nombre de baies Acidité, Sucres,
Nombre de pépins,
Poids frais total des pépins
Poids frais moyen par pépin
% matière sèche des pépins
...



On ne peut cartographier que ce qui varie (caractère phénotypique => choix des populations) ou que ce qui est polymorphe (marqueurs => problème des cartes incomplètes)

QTLs de teneur en anthocyanes détectés dans la population Syrah X Grenache (thèse A. Fournier-Level)

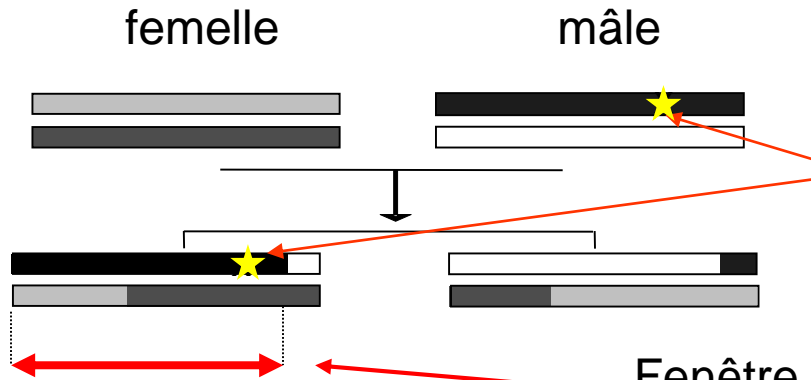


➔ On ne peut détecter que ce qui varie chez les deux parents du croisement = une petite partie de la variabilité existante

Deux forces en action à l'échelle des populations qui expliquent le polymorphisme des génomes des individus qui les composent: **la recombinaison et la mutation**

La génétique d'association

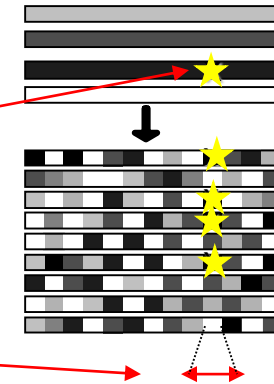
Population en ségrégation



Allele à un site
« fonctionnel »

Fenêtre de déséquilibre de
liaison due aux événements de
recombinaison

Populations naturelles



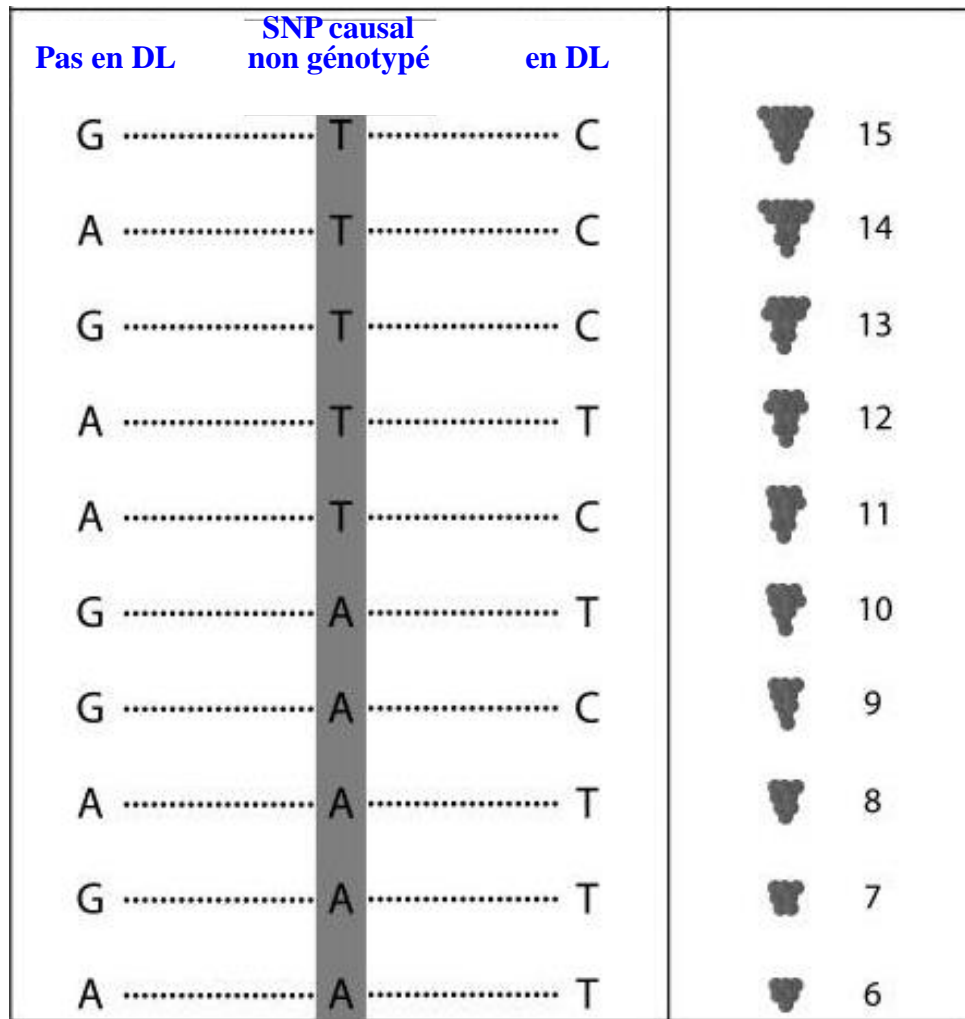
→ Intervalle de confiance large
autour du QTL

→ Si une association significative est détectée, il
y a une plus forte probabilité qu'elle soit proche
du site fonctionnel

→ Plus forte résolution pour l'association caractère/marqueur en
génétique d'association

Autre intérêt de travailler sur des populations naturelles?
Notion de population naturelles ici

Le déséquilibre de liaison



Génétique d'association

P: Variation
phénotypique



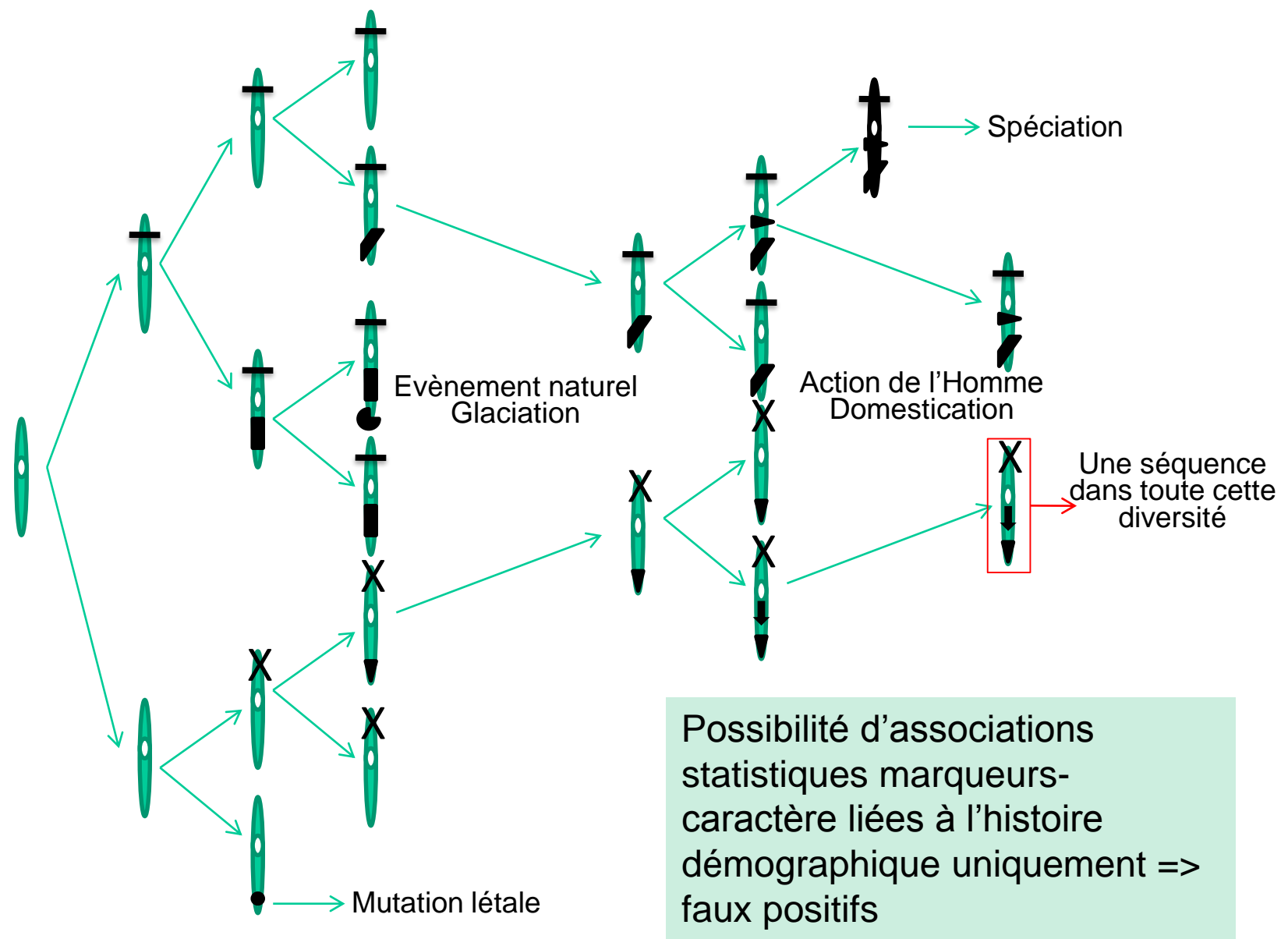
G: Variation génotypique

	Marqueur1	Marqueur2	Marqueur3
ind1	A:A	T:T	T:T
ind2	G:A	T:C	T:C
ind3	G:G	T:C	T:T
ind4	G:G	T:C	T:C
ind5	G:G	T:T	T:C

Modèle statistique:
 $P = G + \epsilon$

Coefficient de corrélation
et la probabilité associée

Dans les populations naturelles: le polymorphisme observé actuellement est le résultat d'histoires démographiques anciennes

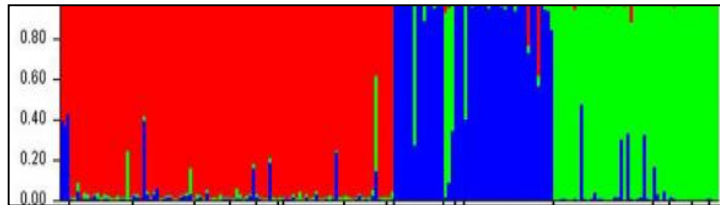


Génétique d'association

G: Variation génotypique

	Marqueur1	Marqueur2	Marqueur3
ind1	A:A	T:T	T:T
ind2	G:A	T:C	T:C
ind3	G:G	T:C	T:T
ind4	G:G	T:C	T:C
ind5	G:G	T:T	T:C

Q: Structure de la population



P: Variation
phénotypique



K: Matrice d'apparentement
des individus 2 à 2

	ind1	ind2	ind3	ind4	ind5	ind6
ind1	_	0,124	0,158	0,149	0,579	0,254
ind2	0,124	_	0,258	0,269	0,697	0,655
ind3	0,158	0,258	_	0,548	0,215	0,156
ind4	0,149	0,269	0,548	_	0,157	0,559
ind5	0,579	0,697	0,215	0,157	_	0,156
ind6	0,254	0,655	0,156	0,559	0,156	_

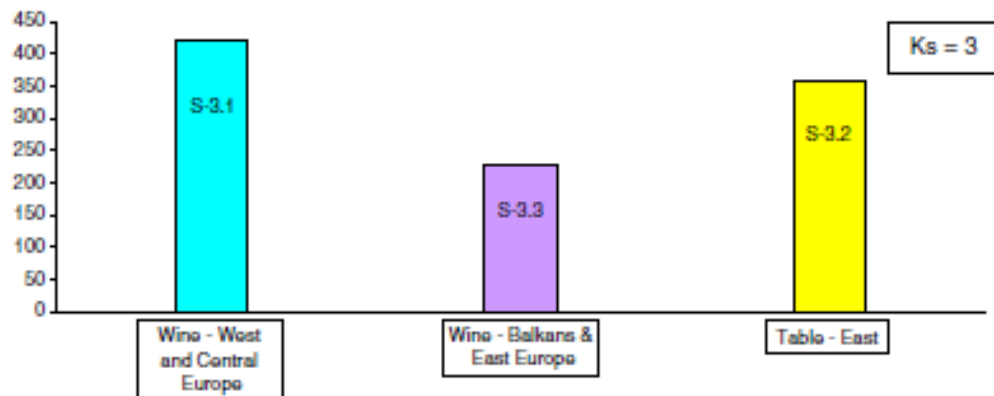
Modèle statistique:
 $P = G + Q + K + \epsilon$

Coefficient de corrélation
et la P-value associée

Structuration des ressources génétiques chez la vigne cultivée sur la base de marqueurs moléculaires

Génotypage de 4000 accessions de vigne cultivée (*Vitis vinifera*) avec 20 SSR

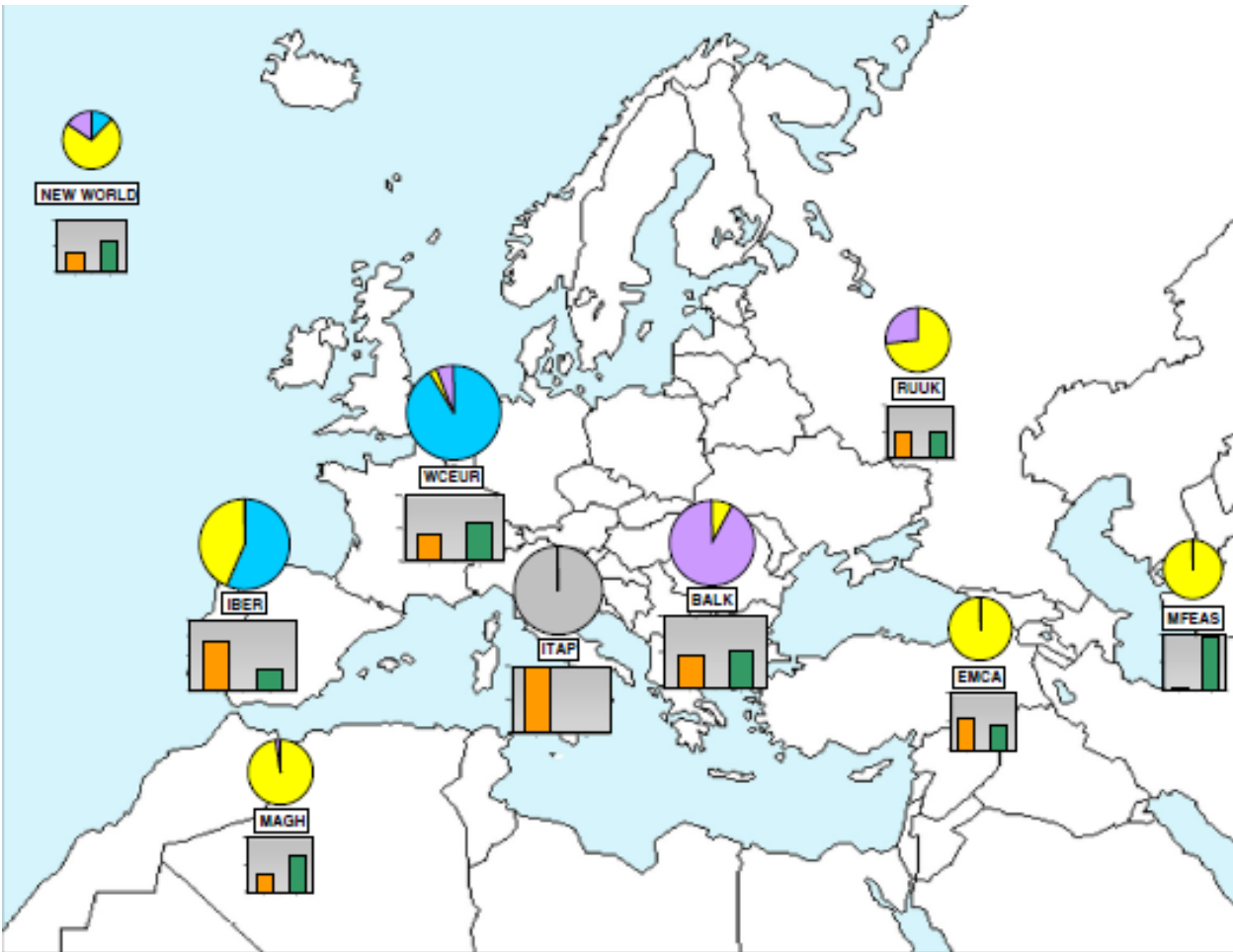
→ 2261 (2096) génotypes non redondants



+ deux sous-populations: péninsule Ibérique & Maghreb et tables récents d'Italie & eur. centrale

3 sous-populations : Cuves Eur. ouest et centrale; cuve Eur Est et Balkans et table Eur est, Asie, caucase...

Structuration des ressources génétiques chez la vigne cultivée sur la base de marqueurs moléculaires

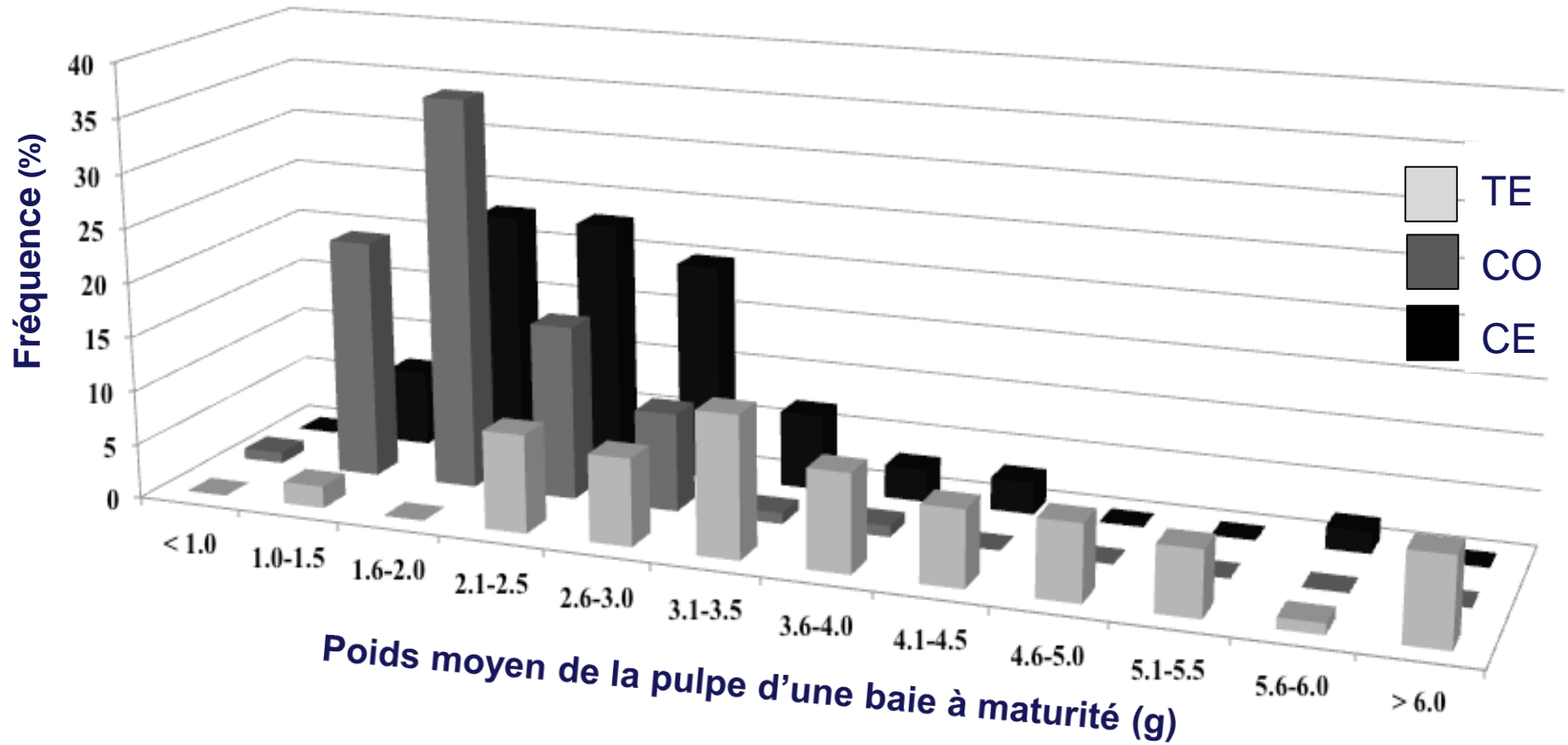


- 3 sous-populations
- Cuve OUEST (●)
 - Cuve EST + Centre (●)
 - Table EST (●)

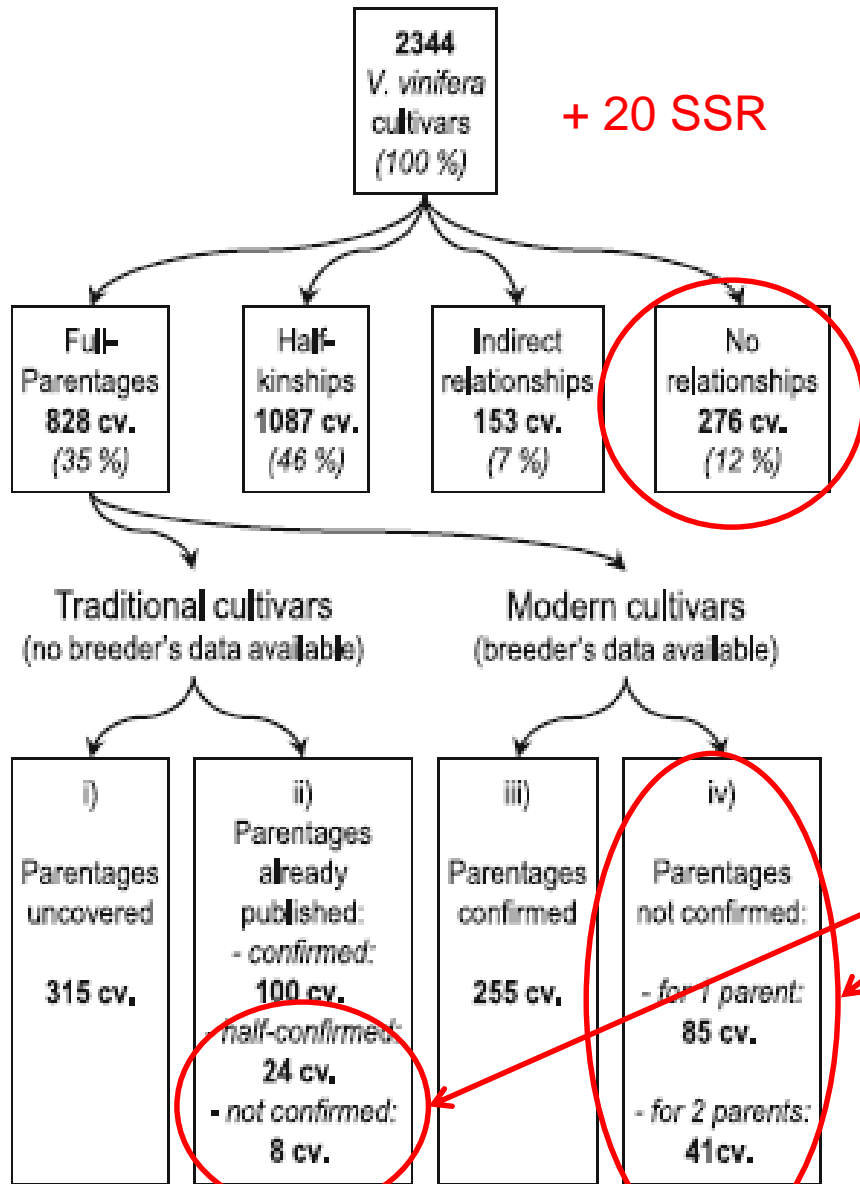
- Structure génétique des individus des populations:
- Avec admixture (■)
 - Sans admixture (■)

Si on ne tient pas compte de cette structure, on risque de détecter de fausses associations

Variation du poids de la baie dans chaque sous-population



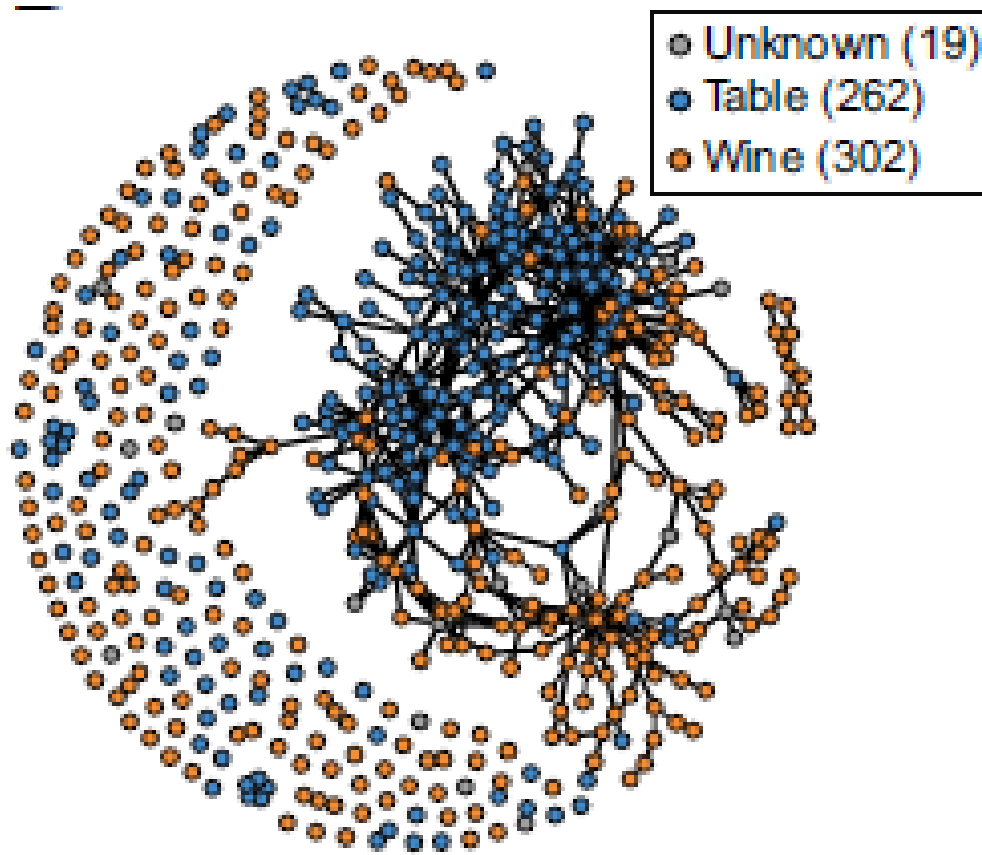
Relation de parentées chez la vigne cultivée



12% seulement des cultivars de la collection de Vassal n'ont aucune relation de parenté détectable

De nombreuses relation de parenté publiées fausses

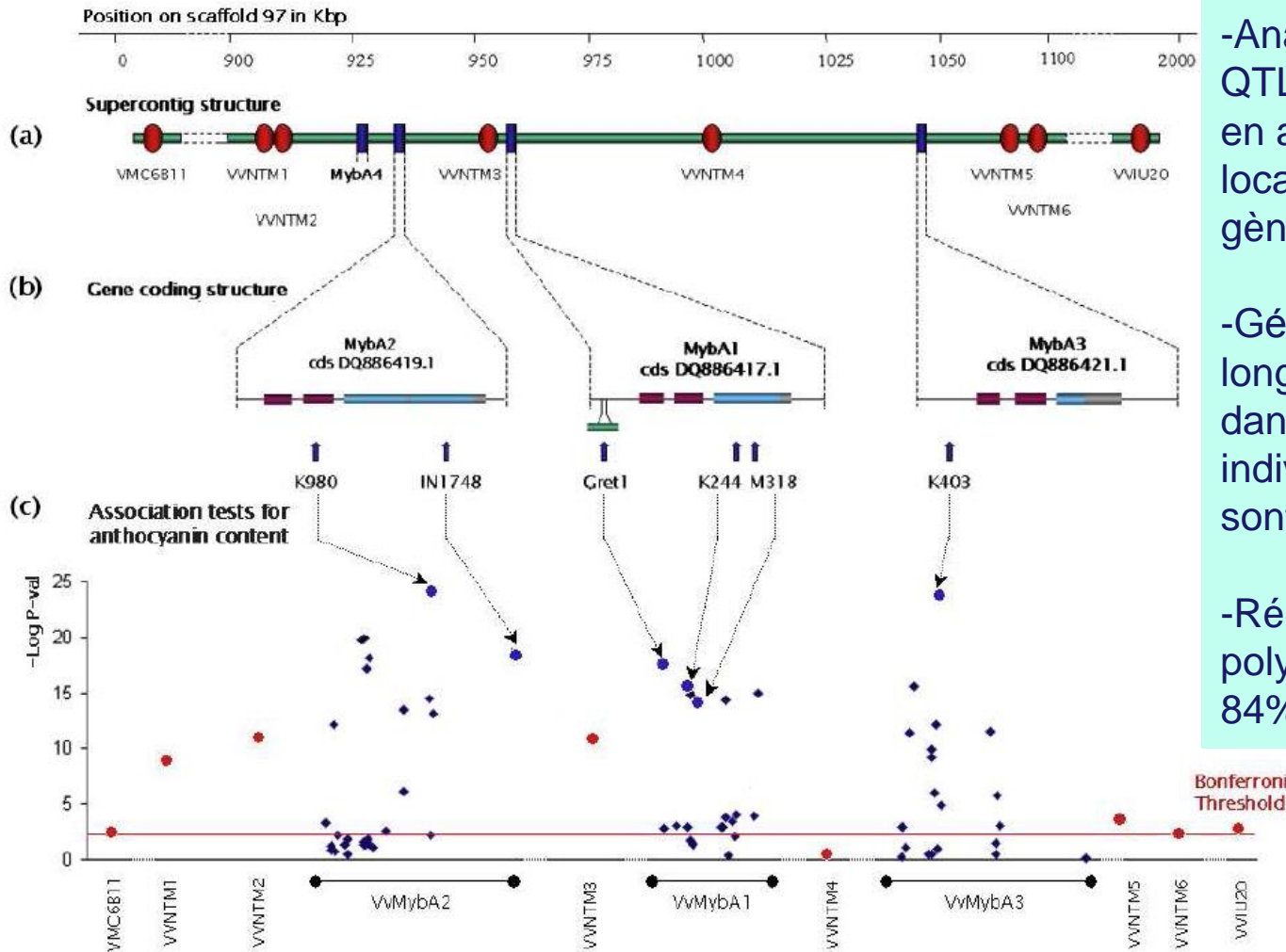
Relation de parentés chez la vigne cultivée



Relations de parenté observées dans 583 cultivars uniques à l'aide de 5387 marqueurs SNP

1 nœud contient 58% des cultivars (384)

Combinaison des deux approches et déterminisme génétique de la variation quantitative de la couleur de la baie (Fournier-Level et al Genetics 2009)



-Analyse QTL classique : 1 QTL explique 62% du contenu en anthocyanes des baies. Co-localise avec le cluster de gènes Myb.

-Génétique d'association le long du cluster de gènes Myb dans une population de 141 individus : 32 polymorphismes sont associés significativement

-Régression multiple: 5 polymorphismes expliquent 84% de la variation observée

28 gènes candidats pour la variation de la taille de la baie (Thèse C. Houel)

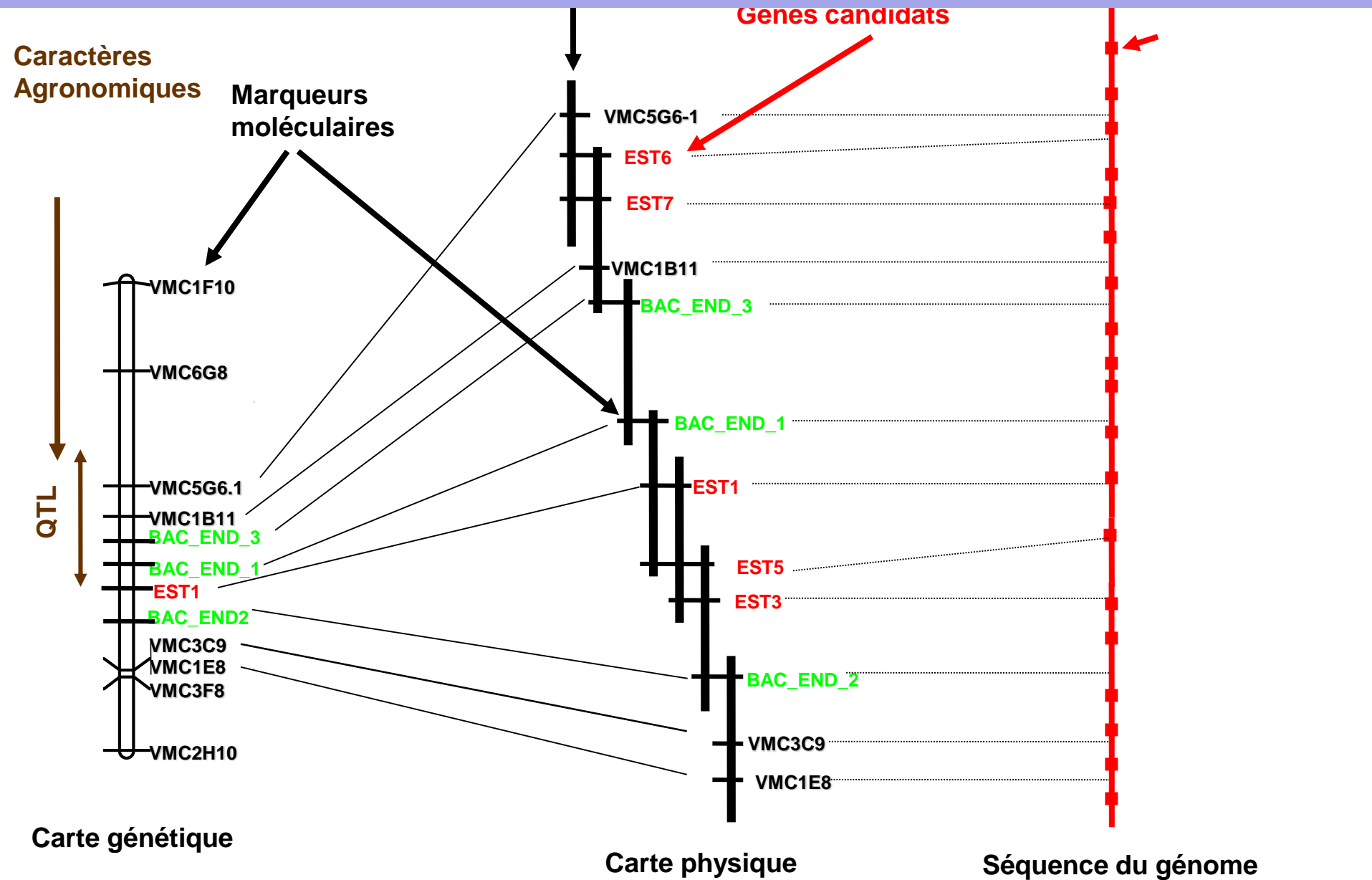
Nom du gène	Gène orthologue paternel (espèce)
VvAGL11	<i>Agamous like 11</i> (Arabidopsis)
VvPMEI2	<i>MMI9.1</i> (Arabidopsis)
VvSPY	<i>SPINDLY</i> (Arabidopsis)
VvRPL24	<i>Ribosomal protein like 24</i> (Arabidopsis)
VvFw2.2.2	
VvFw2.2.1	<i>Fw2.2</i> (Tomate)
VvFw2.2.3	
VvSUN	<i>SUN</i> (Tomate)
VvOvate1	
VvOvate2	<i>Ovate</i> (Tomate)
VvHMGR1	
VvHMGR2	<i>HMG1</i> (Arabidopsis)
VvHMGR3	
VvAP3	<i>Apetala 3</i> (Arabidopsis)
VvAGL22.2	<i>Agamous like 22</i> (Arabidopsis)
VvAGL17.1	<i>Agamous like 17</i> (Arabidopsis)
VvAGL12	<i>Agamous like 12</i> (Arabidopsis)
VvAGL8	<i>Agamous like 8 - Fruitfull</i> (Arabidopsis)
VvMADS1	<i>Agamous</i> (Arabidopsis)
VvMADS6	<i>Apetala 1</i> (Arabidopsis) [§]
VvYABBY2	<i>Fasciated</i> (Tomato)
VvMADS9	<i>Pistilata</i> (Arabidopsis)
VvJMJ	<i>F2K11.14</i> (Arabidopsis)
VvMADS4	<i>Agamous like 9 - Sepallata 3</i> (Arabidopsis)
VvHB13	<i>Homeo Box 13</i> (Arabidopsis)
VvSVP2	–
VvBURP1	<i>Unknown seed protein like 1</i> (Arabidopsis)
VvHD2C	<i>HD2C</i> (Arabidopsis)

- 4: QTL taille de la baie (Mejia et al, 2011)
- 17: bibliographie (Tomate, Arabidopsis) et/ou expression ovaire et baie de vigne
- 7: banque d'ADNc soustractive – puce entre l'Ugni Blanc mutant sans pulpe & sauvage

Gènes présentant des SNP significativement associés à des paramètres de variation de la taille de la baie

Nom du gène	Type de SNP	Localisation	P-value*	Paramètres
<i>VvJMJ</i>	[T/A]	Exon	9,52E-05	Longueur de la baie
			2,51E-05	Poids de la baie
<i>VvHMGR1</i>	[T/C]	Exon	1,56E-05	Longueur de la baie
	[G/A]	Intron	1,87E-05	Longueur de la baie
<i>VvAGL11</i>	[C/T]	Intron	9,22E-07	Nombre de pépins
	[G/T]	Intron	3,07E-06	Nombre de pépins

Comment aller vers l'identification d'un polymorphisme causal?



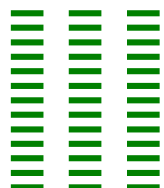
Stratégie de séquençage: Whole Genome Shotgun (WGS)



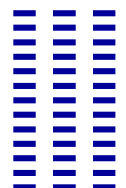
Extraction des chromosomes à partir de jeunes feuilles

Fragmentation mécanique

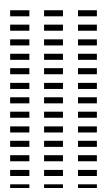
Construction de plusieurs banques de fragments



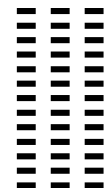
BAC
120kb



Fosmids
40kb

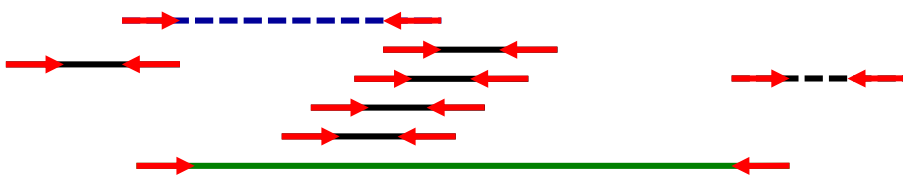


Plasmides
10kb



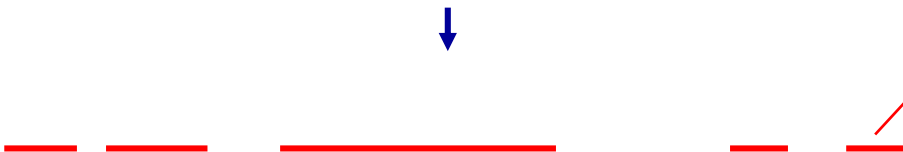
Plasmides
3kb

Séquençage des extrémités de chaque fragment



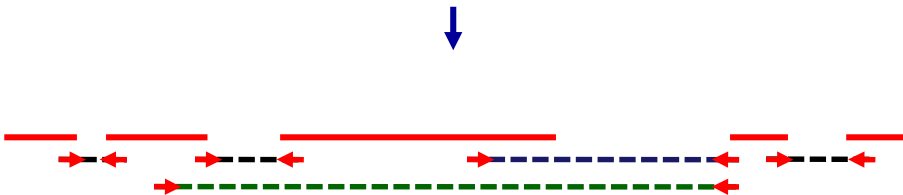
ACGT**AATGGCTTTCCGAAA**
AATGGCTTTCCGAAACGGC
CGACGT**AATGGCTTTCCG**

Assemblage des séquences chevauchantes en *contigs* de séquence



CGACGT**AATGGCTTTCCGAAA**CGGC

Organisation des *contigs* de séquence en *super-contigs*

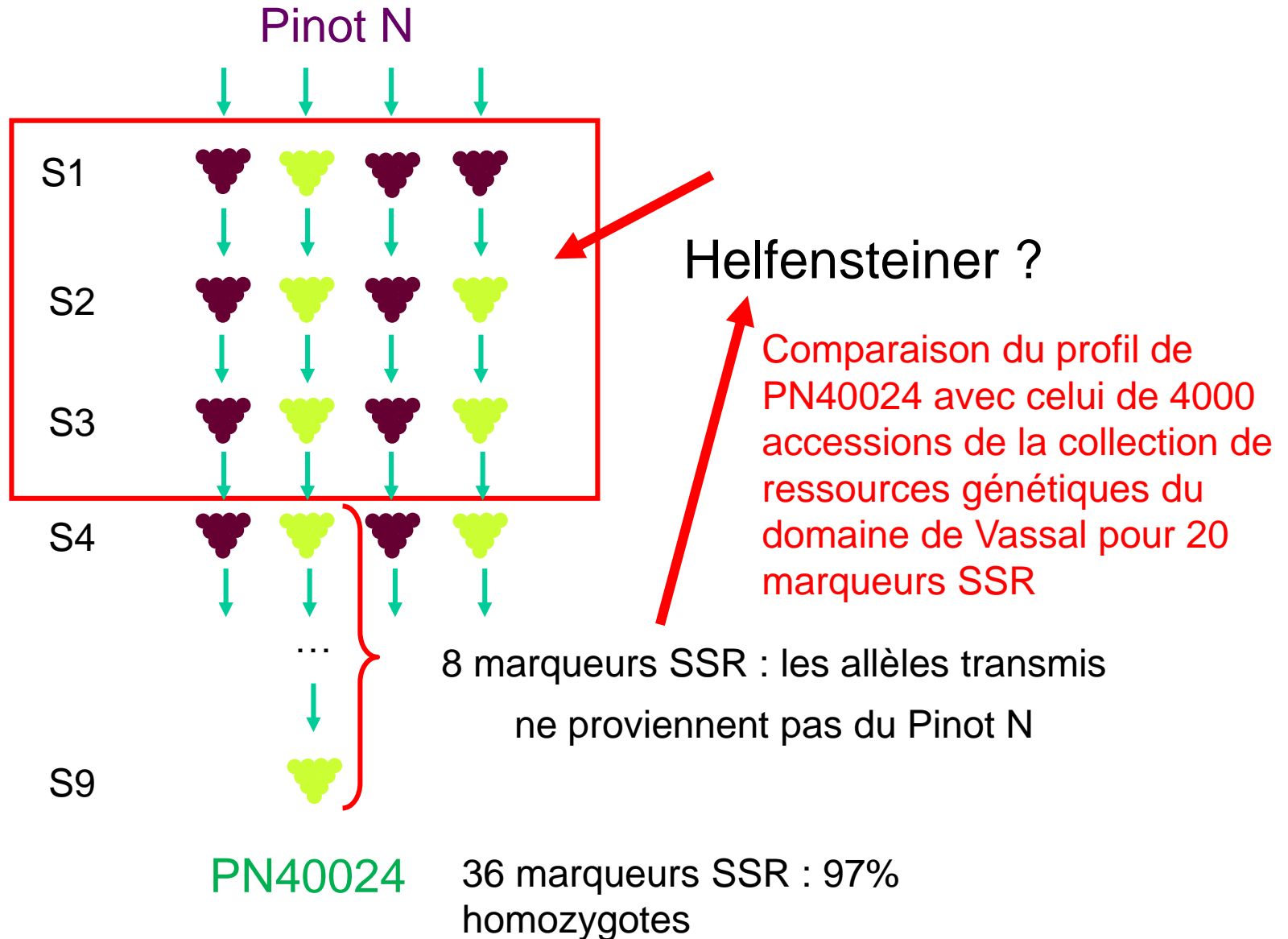


Notion de couverture du génome

Influence de l'hétérozygotie sur une telle stratégie?

Séquençage d'un individu quasi homozygote dérivé du Pinot N par une série d'autofécondations

Bronner et Oliveira 1991



Influence de l'hétérozygotie sur l'assemblage des séquences

- PN40024 (quasi homozygote) Jaillon et al Nature 2007:

6X

- Taille totale de l'assemblage: 487 Mb
- Contigs : Nb= 57 662 N50= 23 Kb

8X

- Taille totale de l'assemblage : 487 Mb
- Contigs : Nb= 19 577 N50= 65.9 Kb

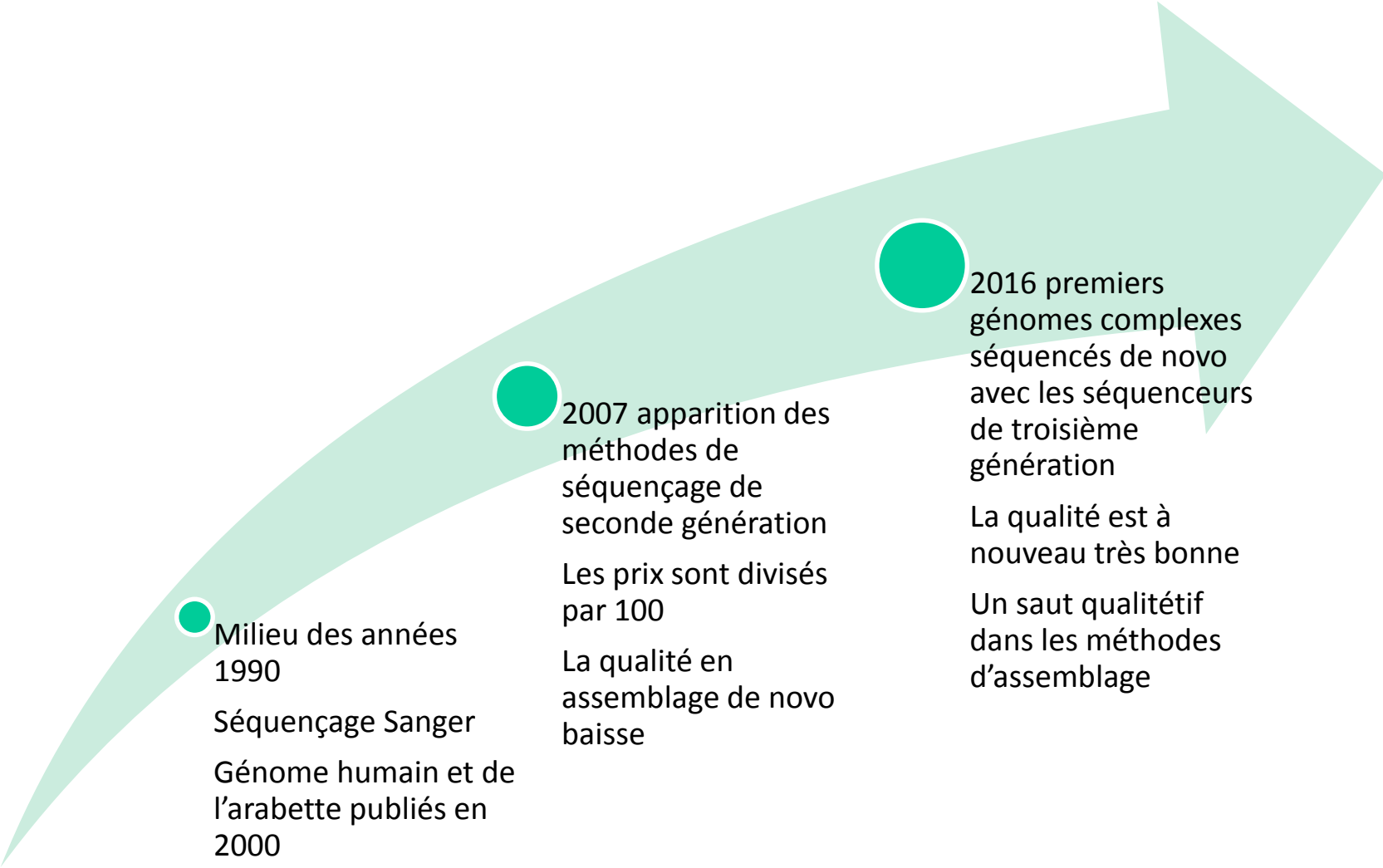
- Pinot Noir (hétérozygote) Velasco et al, PLOSONe 2007:

7X Sanger + 4X pyrosequençage

- Taille totale de l'assemblage: 550 Mb
- Contigs : Nb= 66 165 N50= 10 Kb

Critères de qualité d'un assemblage: longueurs des contigs et des super-contigs

Séquençage des génomes: des révolutions technologiques et informatiques qui s'accélèrent



Milieu des années
1990
Séquençage Sanger
Génome humain et de
l'arabette publiés en
2000

2007 apparition des
méthodes de
séquençage de
seconde génération
Les prix sont divisés
par 100
La qualité en
assemblage de novo
baisse

2016 premiers
génomes complexes
séquencés de novo
avec les séquenceurs
de troisième
génération
La qualité est à
nouveau très bonne
Un saut qualitatif
dans les méthodes
d'assemblage

Séquençage d'ADN par la méthode Sanger: méthode de première génération

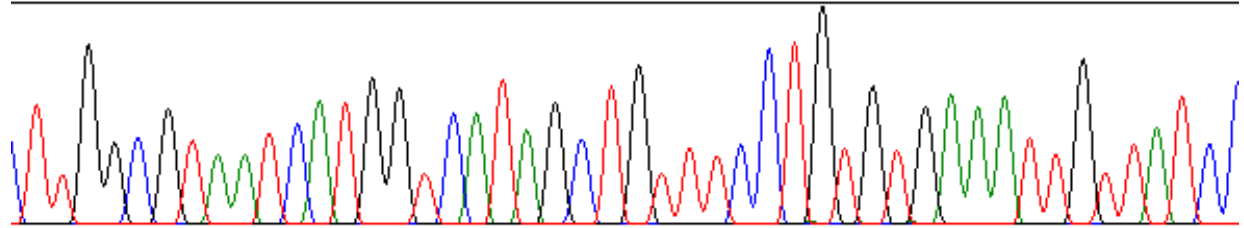
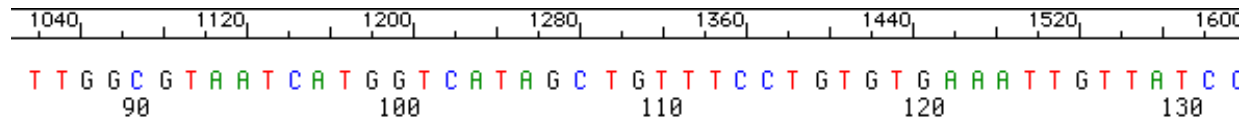
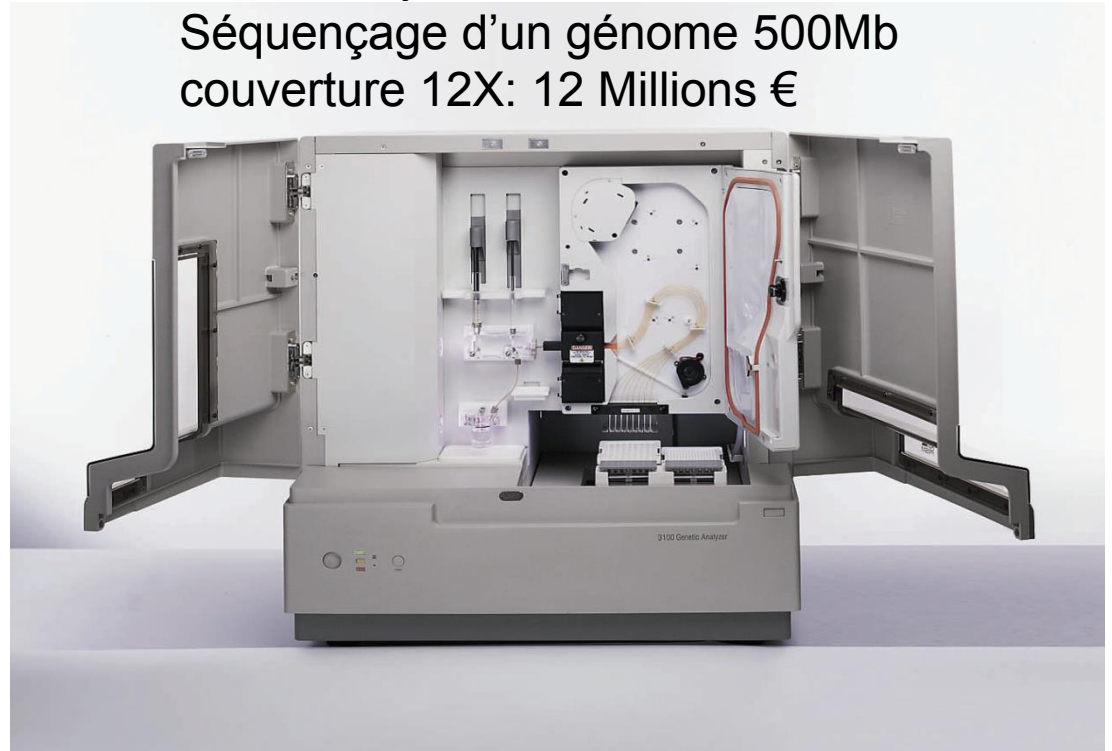
Gel:

-  G
-  T
-  G
-  G
-  A
-  C
-  A
-  T
-  C
-  G
-  C
-  A
-  A
-  C
-  A
-  C
-  C
-  C
-  C
-  T
-  G
-  C
-  G
-  T

GCGAATGCGTCCACACGCTACAGGTG
 GCGAATGCGTCCACACGCTACAGGT
 GCGAATGCGTCCACACGCTACAGG
 GCGAATGCGTCCACACGCTACAG
 GCGAATGCGTCCACACGCTACA
 GCGAATGCGTCCACACGCTAC
 GCGAATGCGTCCACACGCTA
 GCGAATGCGTCCACACGCT
 GCGAATGCGTCCACACG
 GCGAATGCGTCCACAC
 GCGAATGCGTCCACAA
 GCGAATGCGTCCACA
 GCGAATGCGTCCAC
 GCGAATGCGTCCA
 GCGAATGCGTCC
 GCGAATGCGTC
 GCGAATGCGT
 GCGAATGCG
 GCGAATG
 GCGAAT

~11Mb/15 jrs

Séquençage d'un génome 500Mb
 couverture 12X: 12 Millions €



Séquençage Seconde Génération (Next Generation Sequencer)



Fragmentation

En 2007

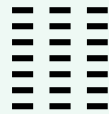
Clonage individuel des fragments dans des bactéries



BAC
120kb



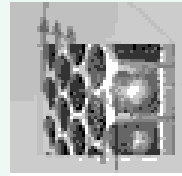
Fosmids
40kb



Plasmids
3kb+10kb

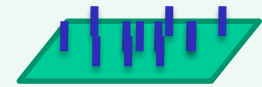
Séquences indépendantes de 500-700 bp en moyenne

Fixation de chaque fragment sur une microbille



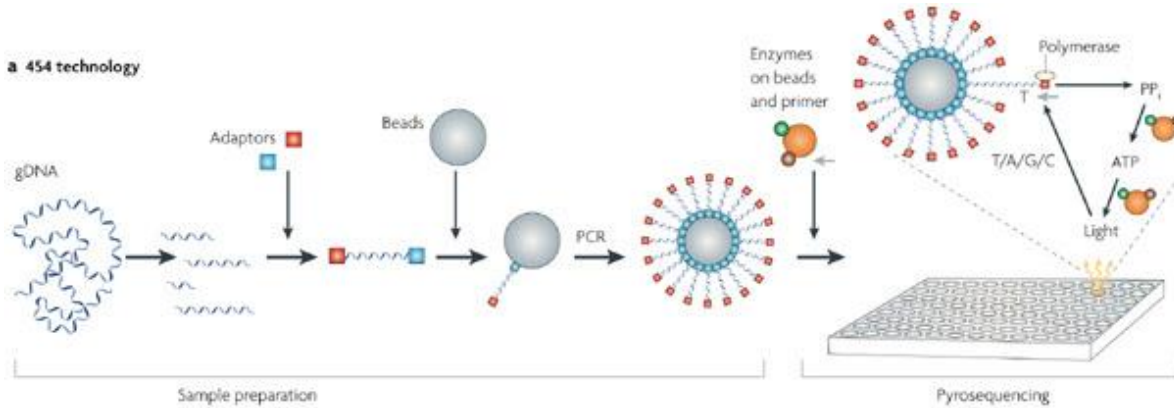
400 000 séquences de 250 bp en moyenne

Fixation de chaque fragment sur une lame



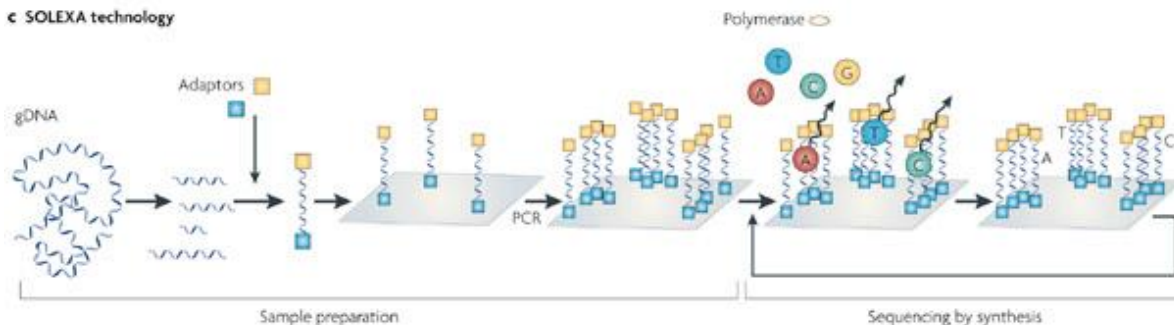
40 millions de séquences de 35 bp en moyenne

Séquençage Seconde Génération (NGS): évolution très rapide



Roche-454: lectures 400 bp
5-6Gb de séquence en 15 jours

Séquençage d'un génome de 500Mb (Vigne par exemple):
15-20X Roche-454 et 50X Illumina-Solexa => ~500k€

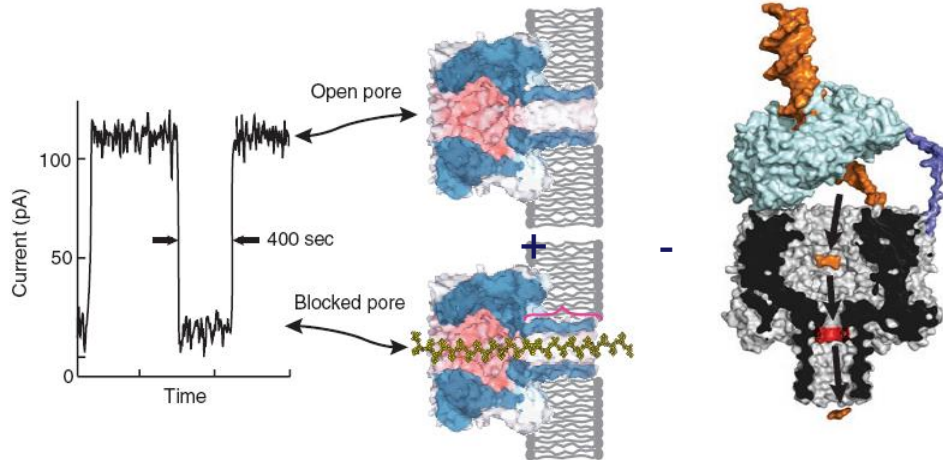


Illumina-Solexa: lectures
100bp des deux extrémités
et 20 Gb de séquence en 15 jours

Séquenceurs de troisième génération: objectif augmenter la longueur des lectures

Séquençage en temps réel = 3^{ème} génération de séquenceurs

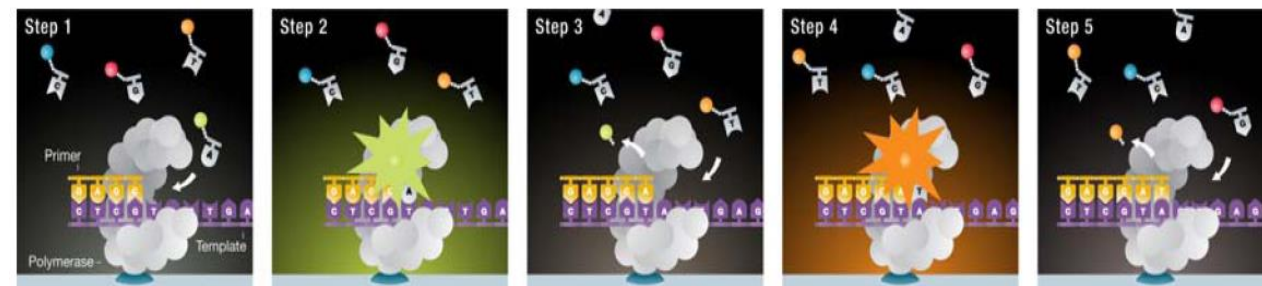
Séquençage Nanopore



- Séquençage de ssDNA
- Longueur de séquence : 50 000 bp
- Génome mammifère < 1000\$ en 24h

Branton et al., 2008 *nat. biotech.*

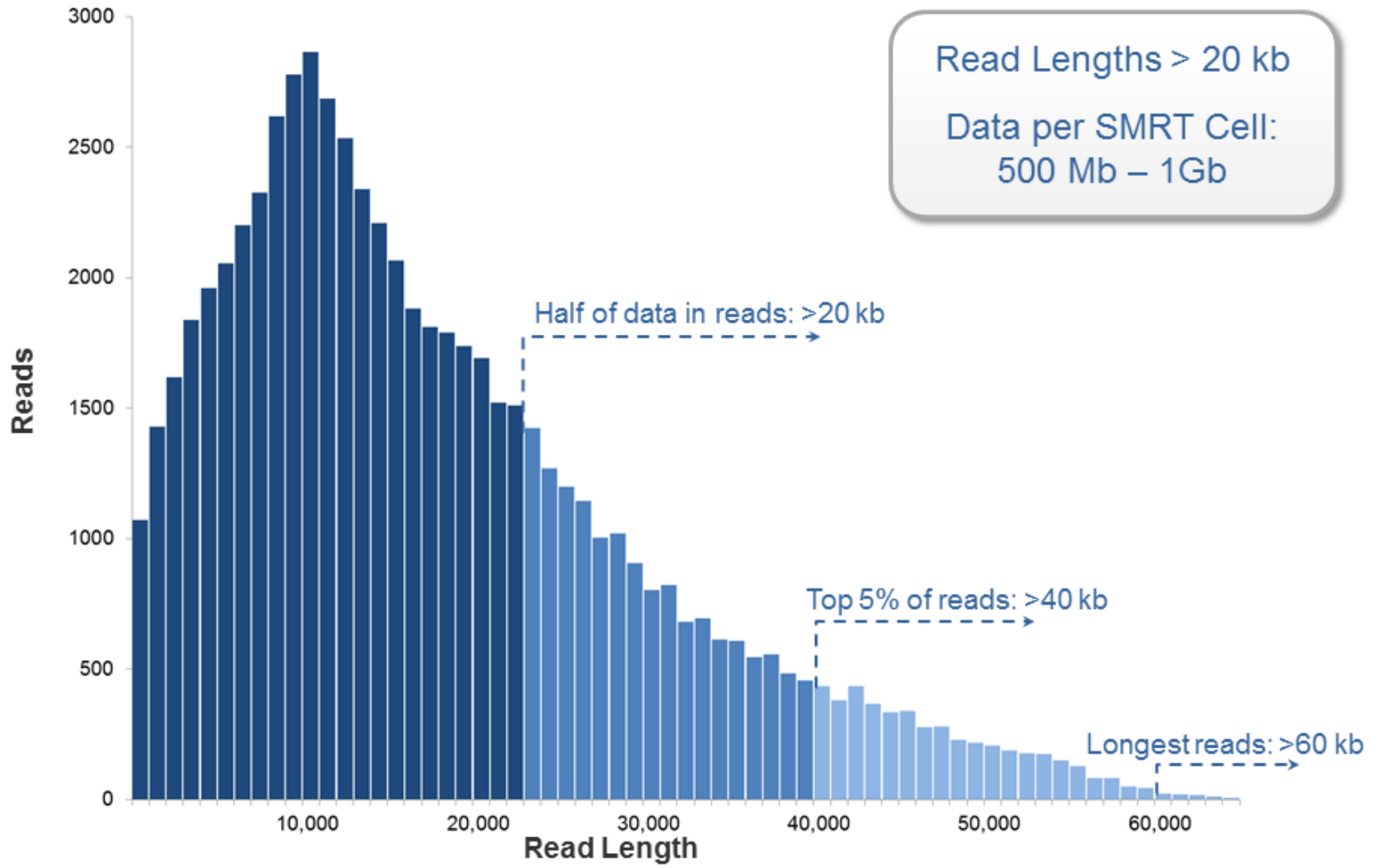
Mesure en temps réel de l'activité de la polymérase - Single Molecule Real Time SMRT™ DNA Sequencing Technology (Pacific Biosciences) :



Polymérase avec fluorophore
nucléotide marqué sur P
Zero-mode waveguides
(détection fluo d'une seule
mol.)

Figure 10. Processive Synthesis with Phospholinked Nucleotides.

Les enjeux majeurs: allonger les séquences, qualité des séquences



Séquenceurs PACBIO

Les enjeux majeurs: améliorer les algorithmes d'assemblage en relation avec les séquences répétées, l'hétérozygotie et la polyploïdie

DeNovoMAGIC™ 2.0

DeNovoMAGIC™ algorithm assembles whole genome seq. data of short reads, into long, phased, and accurate scaffolds.

Input data:

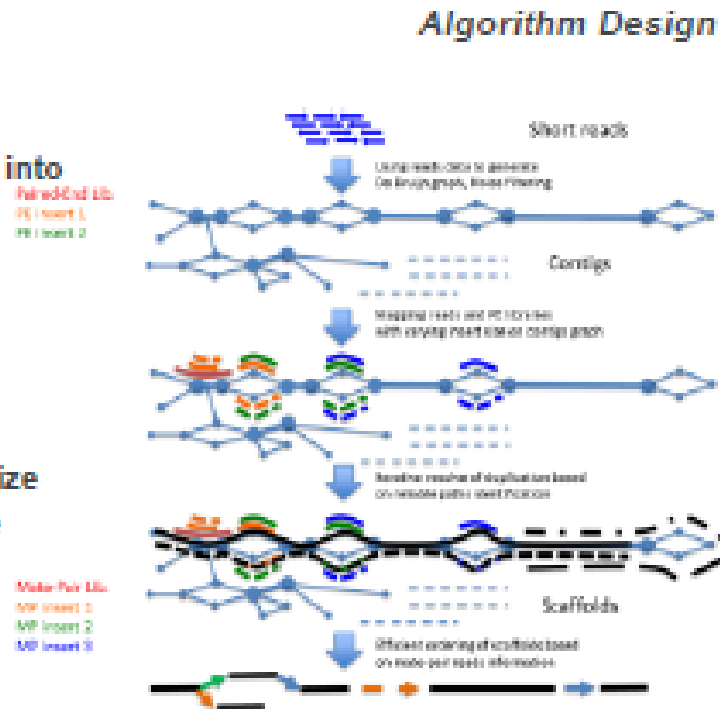
PE (2x250), 450bp insert size

PE (2x160), 800bp insert size

MP (2x160), 3Kb, 5-7Kb, 8-10Kb insert size

Illumina's Chemistry- v4, v2 rapid mode

Total 180-200X coverage.

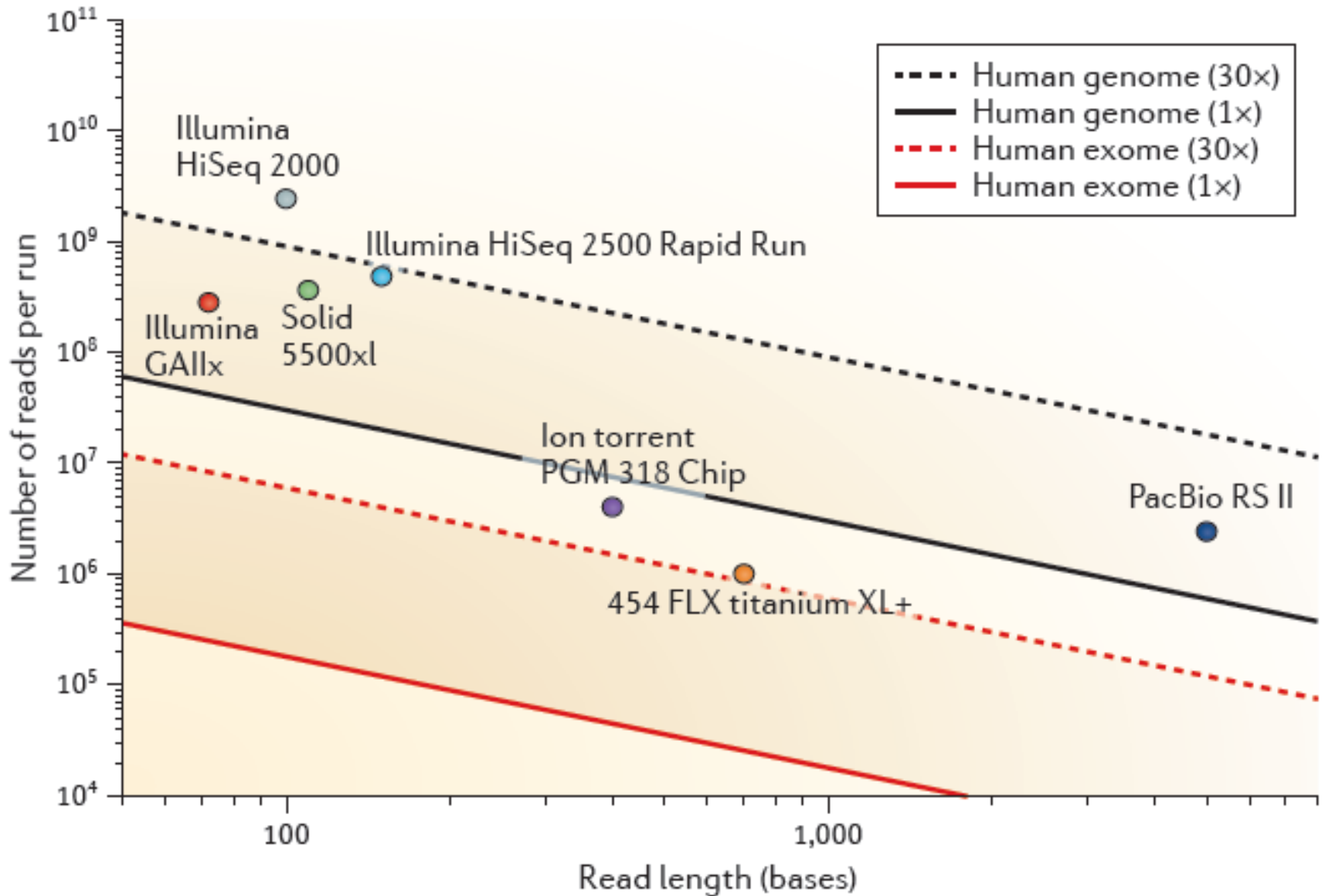


© 2015. All rights reserved to Illumina, Inc. Technology. 2000 141

=> Assemblage de Novo de la séquence du blé!!!!

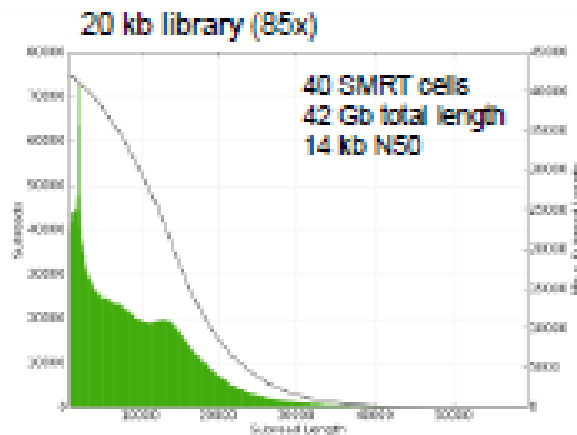
Ronen G (NRGene SA) Plant and Animal Genome Conference XXIV, San Diego, 2016

Les enjeux majeurs: allonger les séquences, qualité des séquences



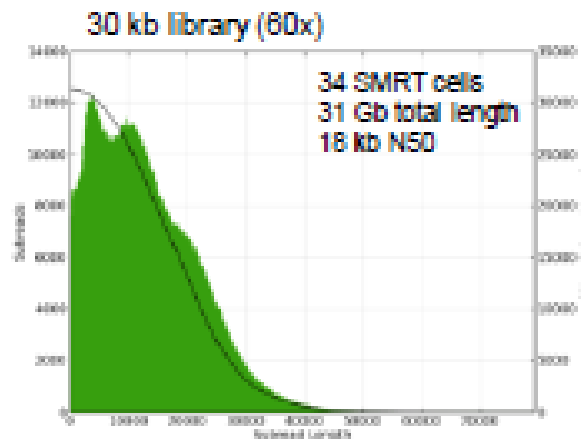
Séquençage du génome du Cabernet Sauvignon en PacBio

- Qualité de l'ADN de départ déterminante
- Cette condition étant remplie: assemblage de très grande qualité d'un génome très hétérozygote



	Falcon-unzip Primary
Assembly Size (Mbp)	805.2
# of Contigs	2,579
Max Contig (Mbp)	3.92
N50 Size (Mbp)	0.436

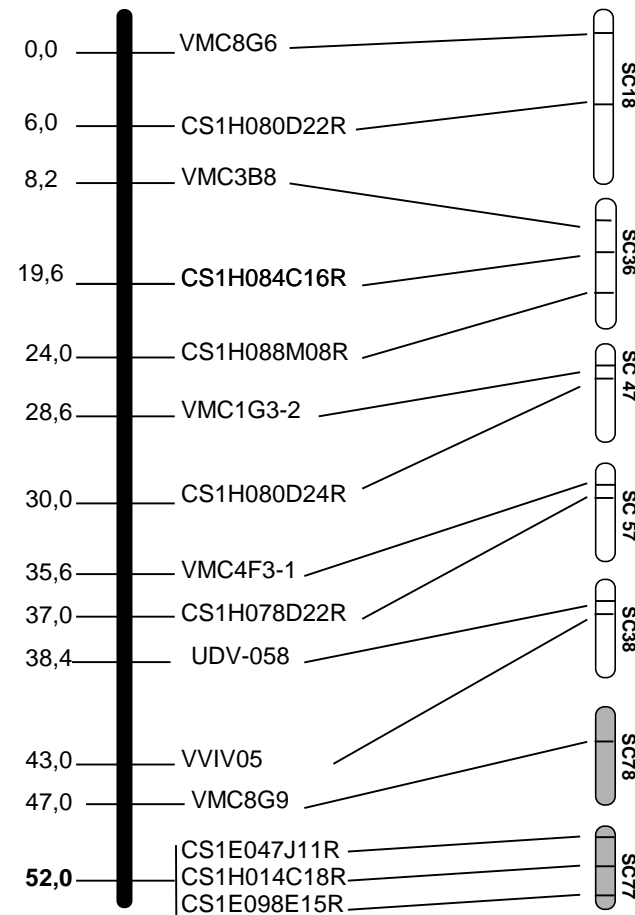
+



	Falcon-unzip Primary
Assembly Size (Mbp)	591
# of Contigs	788
Max Contig (Mbp)	14.0
N50 Size (Mbp)	2.17

Jason Chin (PacBio) - FALCON-unzip (W092)

Ordonnement des super-contigs le long des chromosomes à l'aide d'une carte génétique

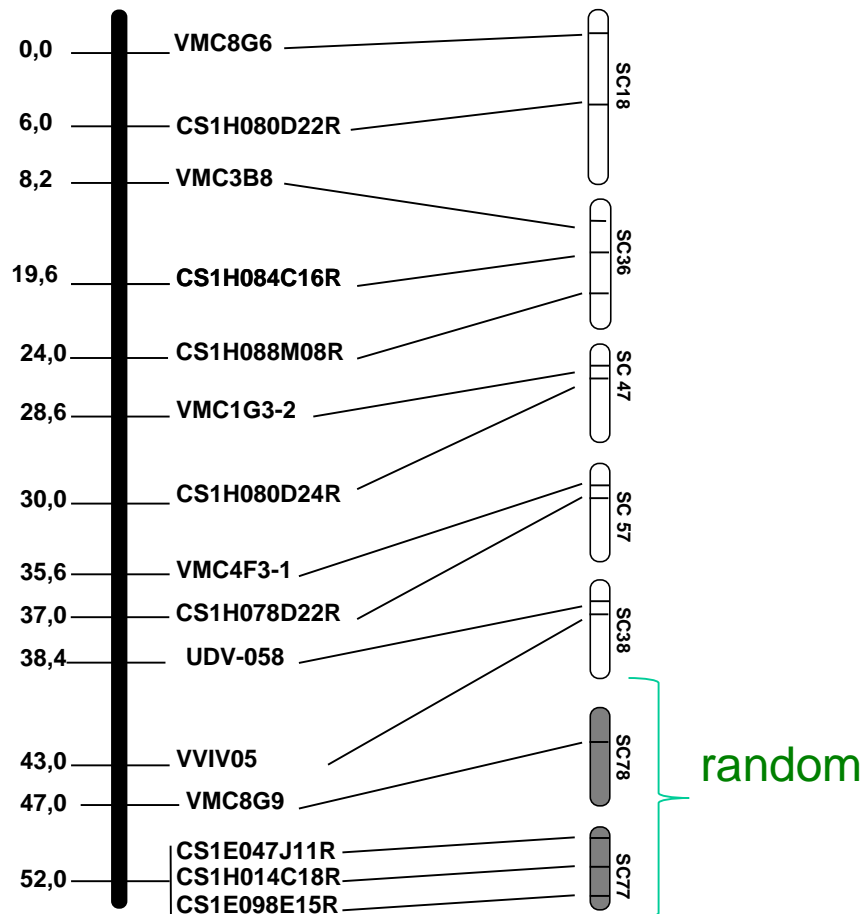


Quels types de marqueurs utiliseriez-vous ? Quelle stratégie mettriez-vous en oeuvre ?

Ordonnancement des super-contigs le long des chromosomes à l'aide d'une carte génétique

Carte génétique:
chromosome 12

Super-contigs de
séquence



Version 8X (carte consensus de 5 petites populations; 409 marqueurs SSR uniques):

- 69% ancrée
- 61% ancrée et orientée

Version 12X (carte consensus de deux grandes populations 2 pop; 514 marqueurs SSR uniques):

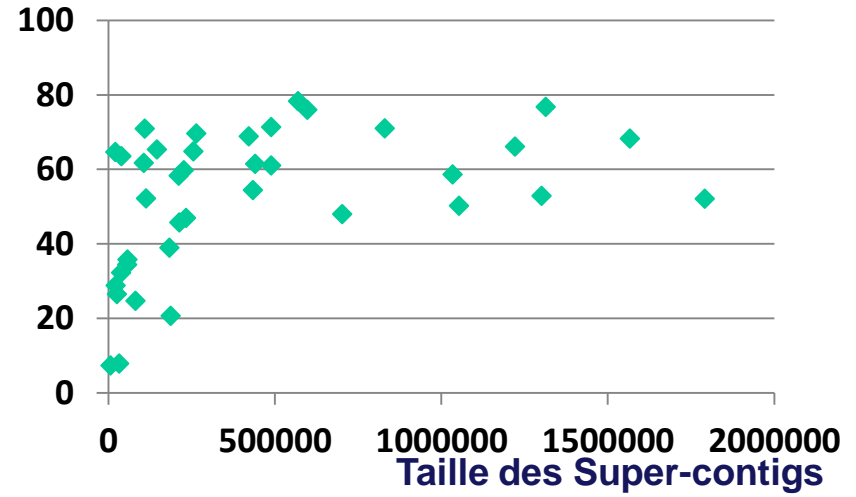
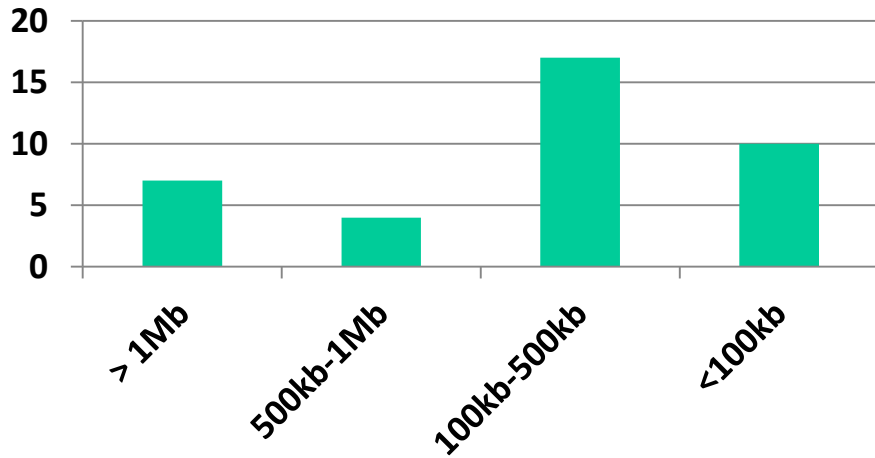
- 85%** ancrée
- 71%** ancrée et orientée

<http://urgi.versailles.inra.fr/>

Caractéristiques des séquences non ancrées et/ou non orientées

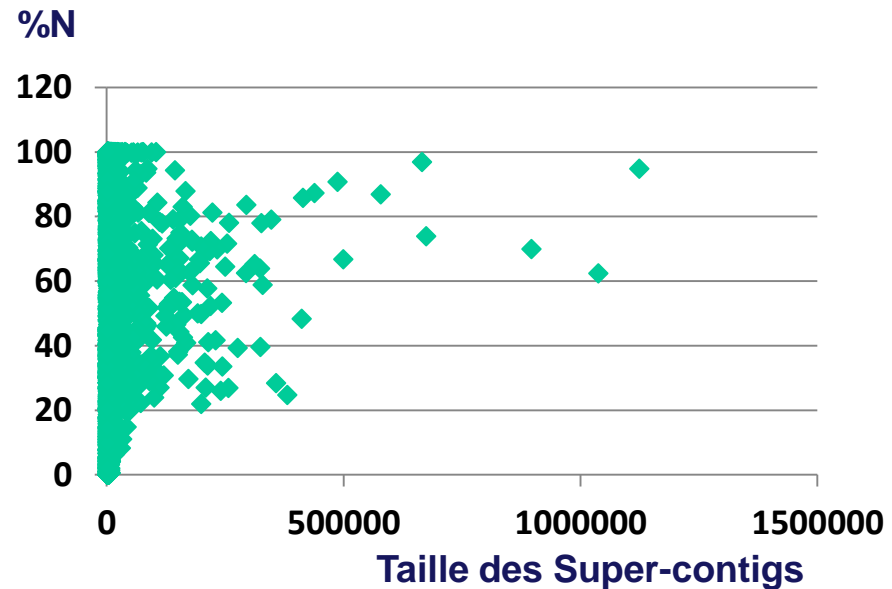
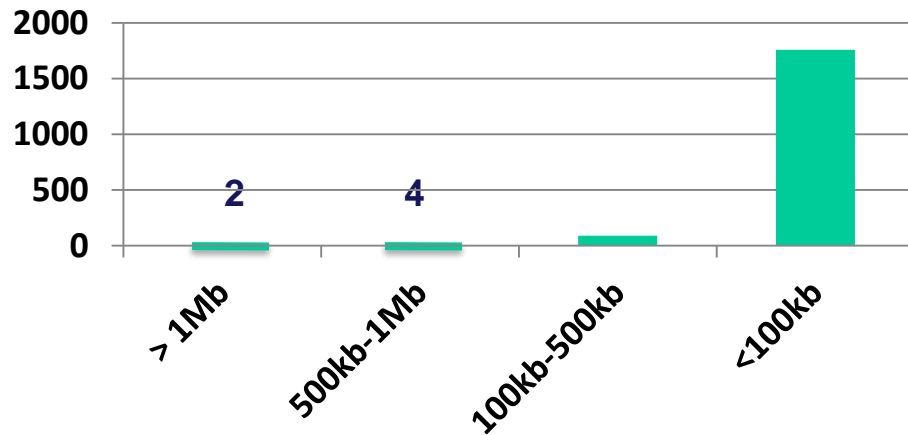
Ancrées sur un chromosome mais non orientées %N

n° super-contigs



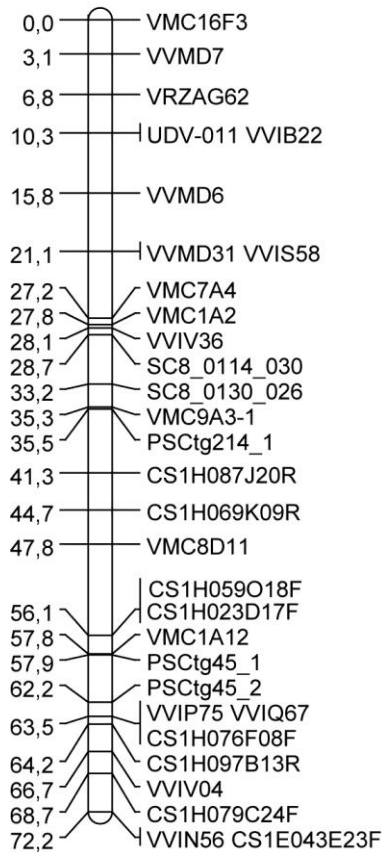
Pas ancrées sur un chromosome

n° super-contigs

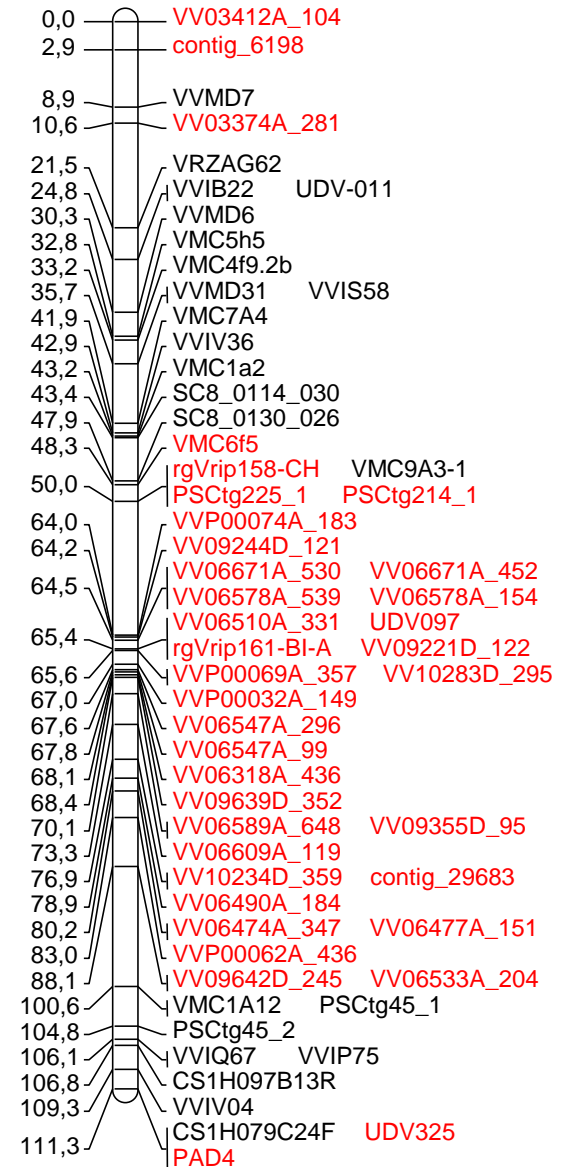


Densification des cartes génétiques

LG7



LG7_ChxBi

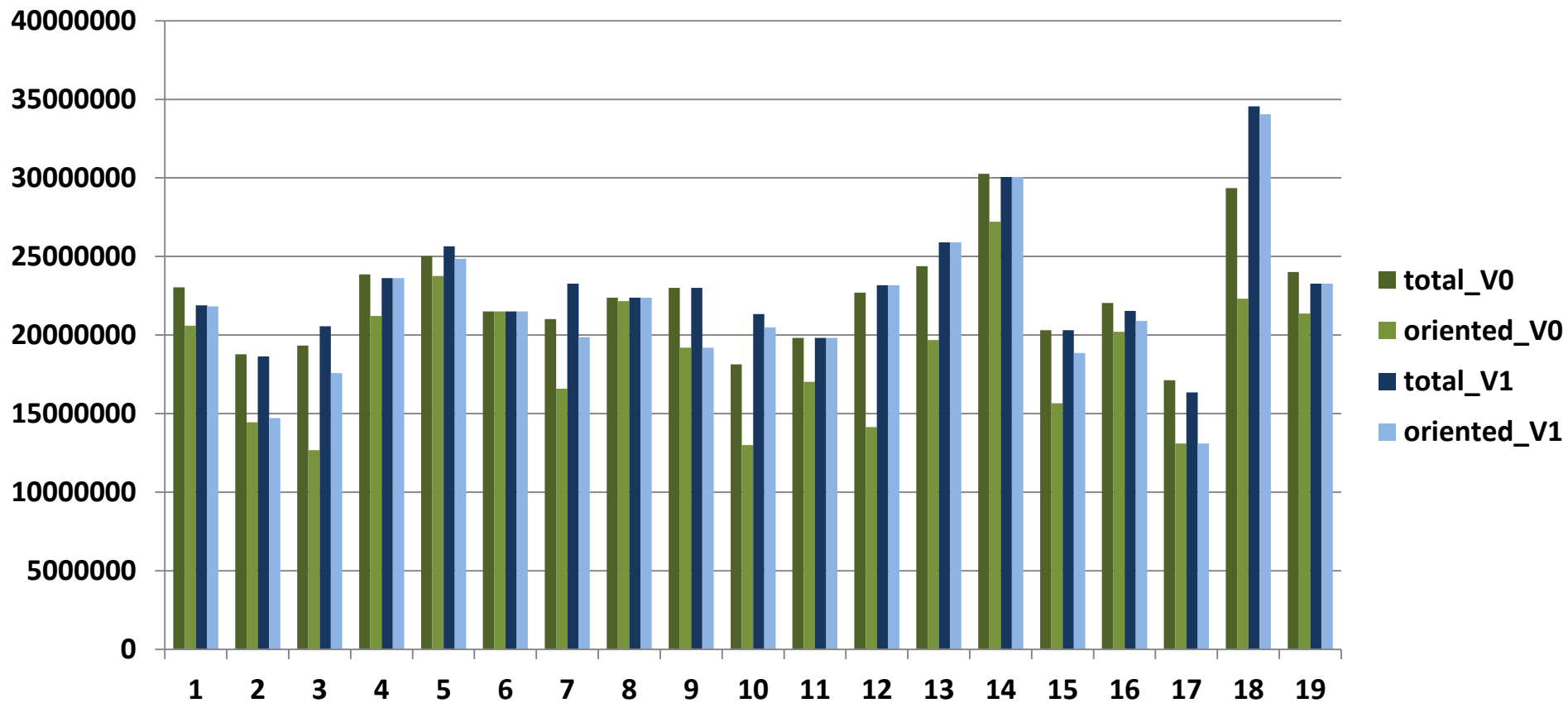


Integrated SxG and ChxBi
(Cipriani et al 2011)

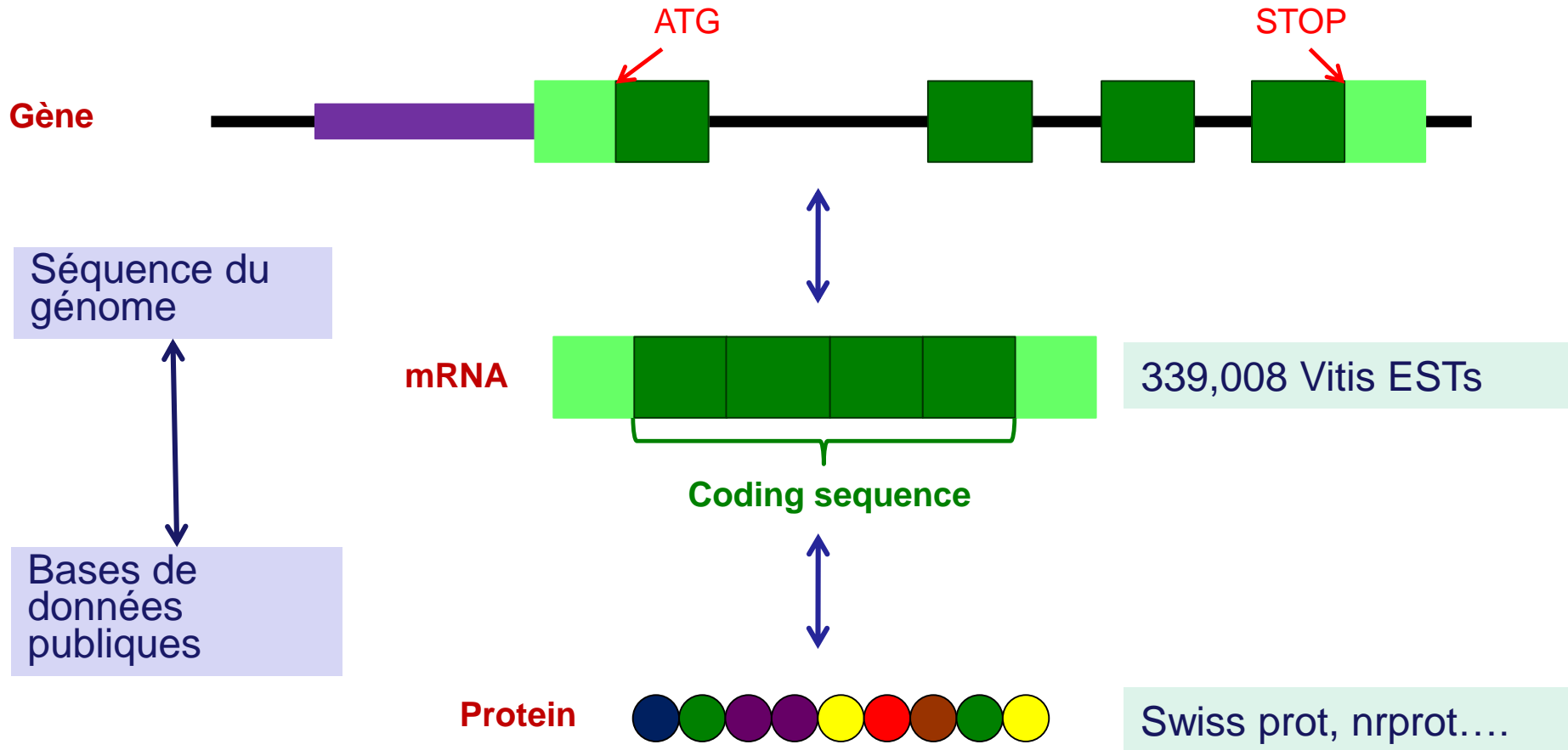
Amélioration de l'assemblage en chromosomes

V0: 426 Mb placée (85% genome) -> V1: 485 Mb placée (97%)

V0: 356 Mb placée et orientée(71% genome) -> V1: 395 Mb placée et orientée (79%)



Annotation : identification de toutes les séquences codantes par annotation automatique



Qu'est-ce qu'un EST?
Comment prédire la fonction d'un gène?

Ressources EST utilisées pour l'annotation automatique de la version 12X du génome de la vigne

1. Viridiplantae (green plants)	:	11,004,234
2. Embryophyta (plants)	:	10,775,868
3. Tracheophyta (vascular plants)	:	10,535,831
4. Spermatophyta (seed plants)	:	10,516,722
5. Magnoliophyta (flowering plants)	:	9,839,606
6. Eudicotyledons	:	5,222,268
7. Vitaceae (grape family)	:	339,012
8. Vitis	:	339,008
9. Vitis vinifera	:	317,348

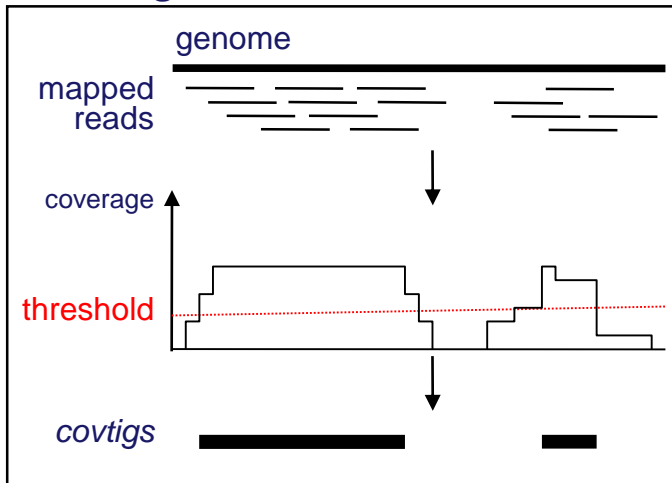
+ cDNA Consortium : 99828 clones **5 banques**

+ RNAseq (Solexa) : ~175 millions reads from 4 tissues **Feuilles, cals, racines, tiges**

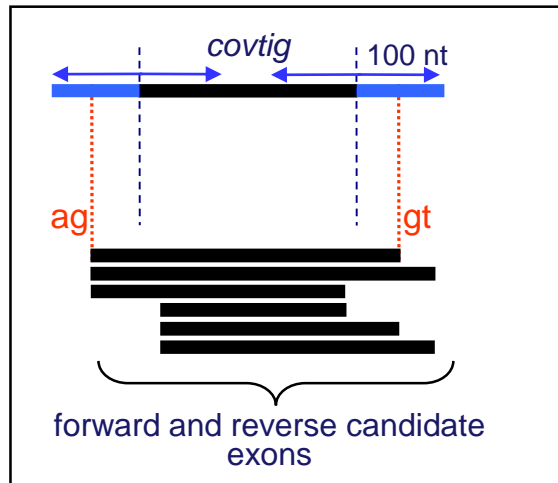
Annotation en utilisant des lectures courtes de séquences EST (Denoeud et al, 2008, Genome Biology)

- 173 millions de lectures Solexa/Illumina GA1 (36 bp)
- Beaucoup s'alignent sur des modèles de gènes
- Mais on trouve aussi des alignements sur des régions non codantes
- Développement d'une méthode pour discriminer le bruit de fond des régions codantes et pour créer des modèles de gènes en connectant des lectures courtes

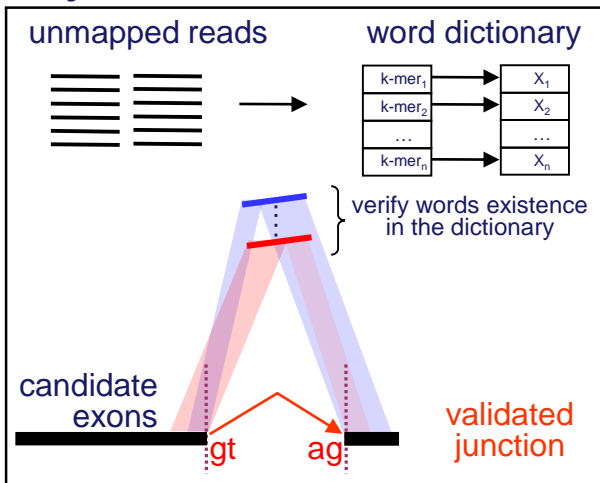
1. *covtigs* construction



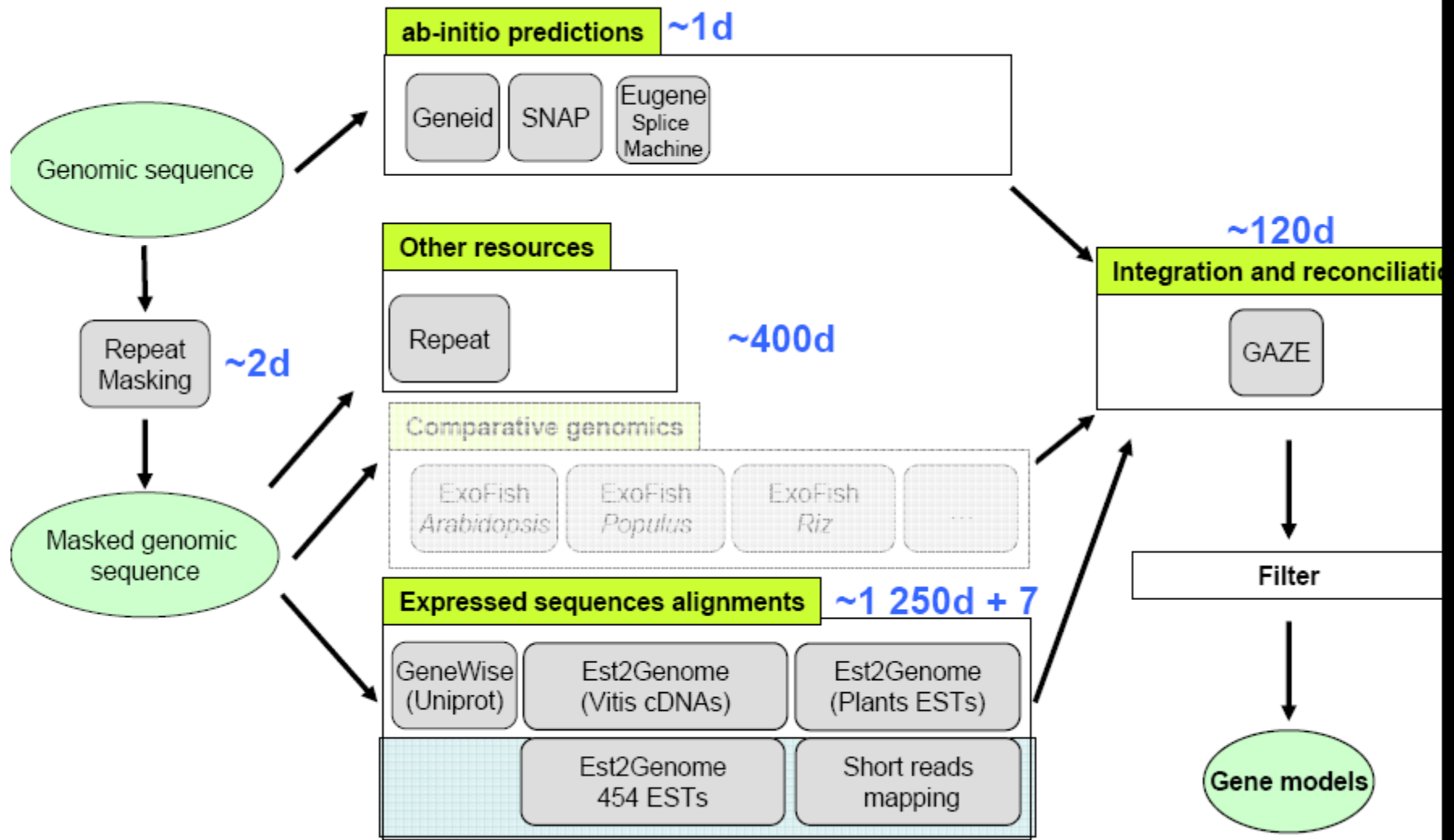
2. candidate exons



3. junction validation



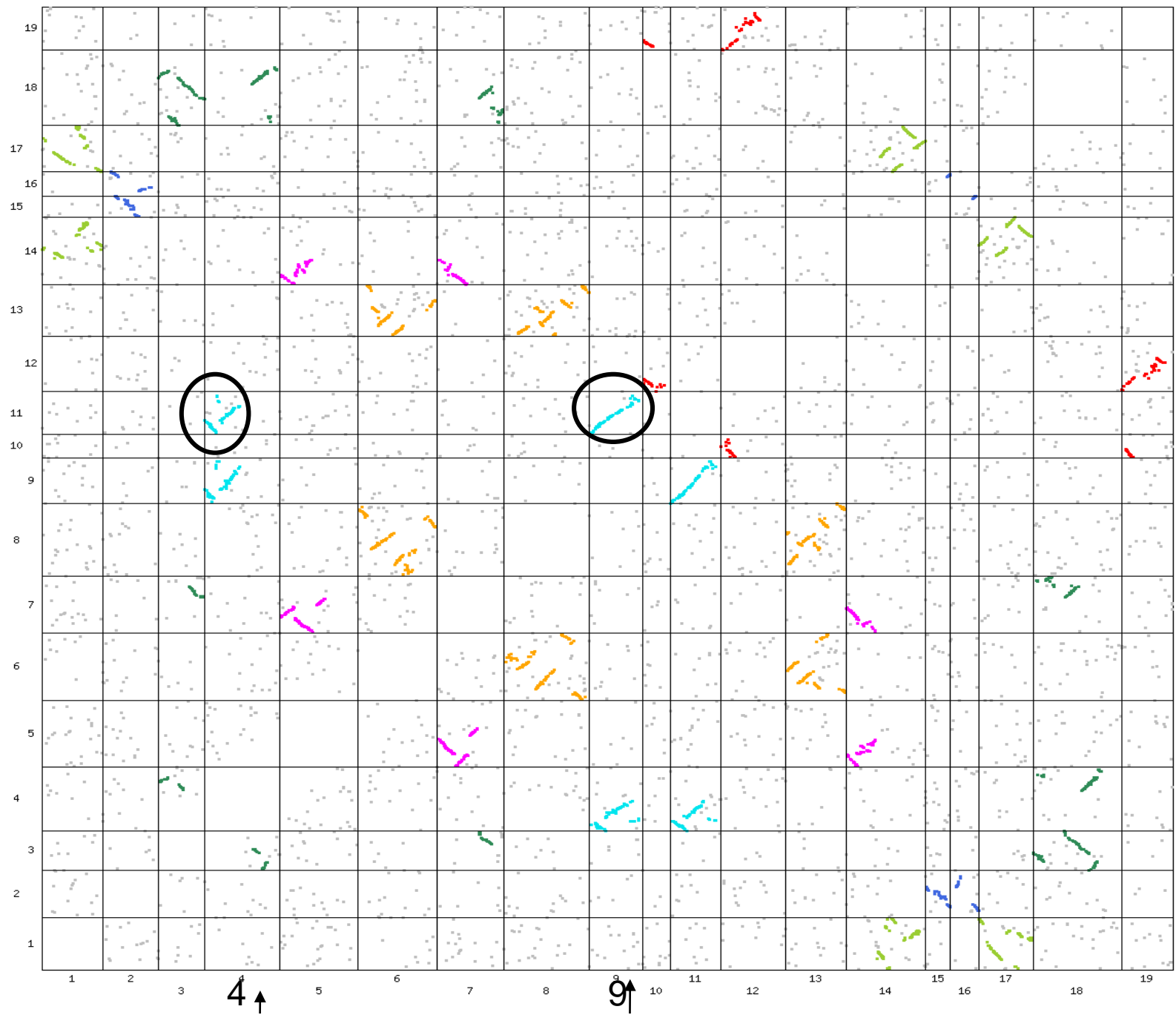
Annotation automatique avec GAZE (Howe et al Genome Res, 2002)



8X 30 434 modèles de gènes
 12X-V0: 26 347 modèles de gènes
 12X-V1: 29 971 modèles de gènes

Une annotation automatique n'est qu'une prédiction qu'il faut valider et améliorer

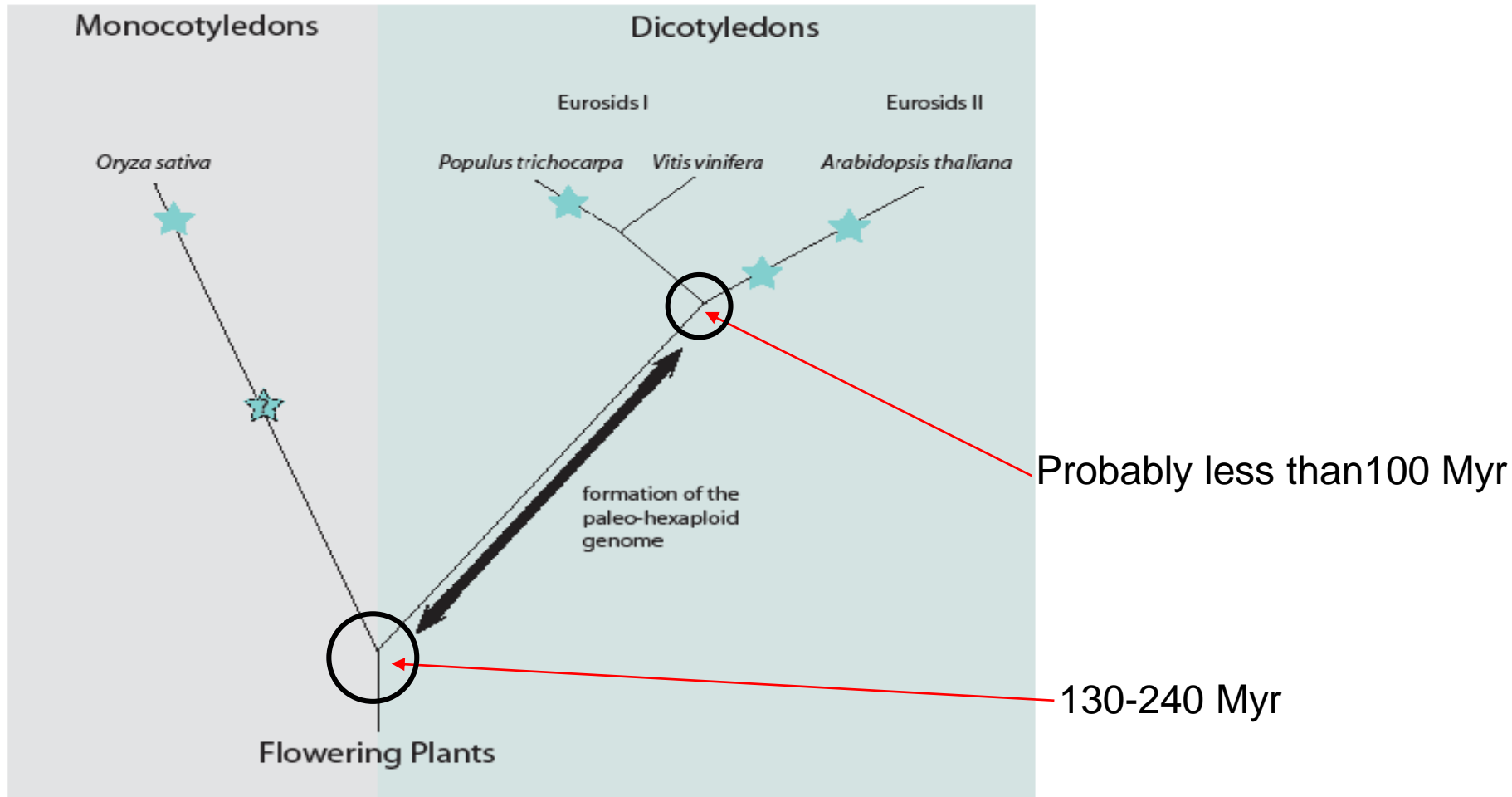
11
→



4
↑

9
↑

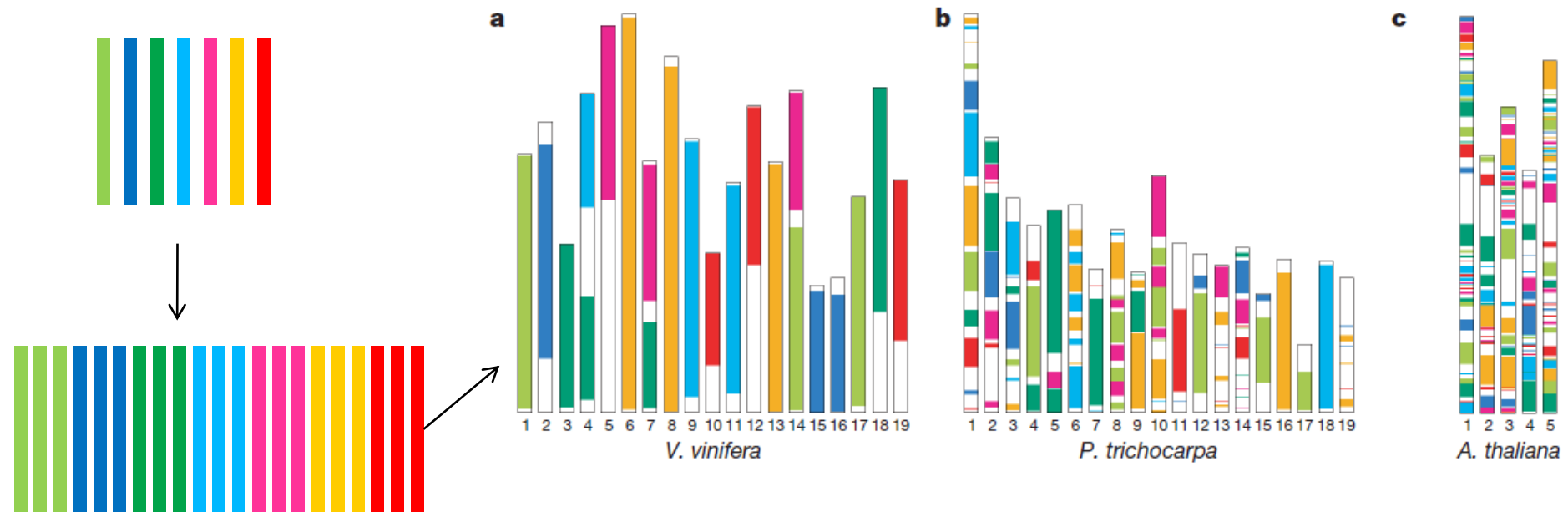
L'ancêtre de la vigne, du peuplier et d'arabidopsis mais pas des monocotylédones avait un génome hexaploïde



Les génomes d'Arabidopsis et du peuplier ont subi une ou deux duplications supplémentaires

Le génome actuel de la vigne a été peu remanié par rapport à celui de son ancêtre hexaploïde

Le génome de la vigne a été peu réarrangé depuis la formation de l'hexaploïde ancestral



Evolution du génome des plantes par:

- Auto ou allo-polypléidisation suivis de réarrangements chromosomiques
- Duplications locales
- perte d'une partie des gènes dupliqués ou évolution différentielle par accumulation de mutations des différentes copies

Annotation expertisée de familles de gènes

De 30% à plus de 50% des gènes qui n'étaient pas détectés par l'annotation automatique
Jusqu'à 50% d'erreurs dans les structures prédites

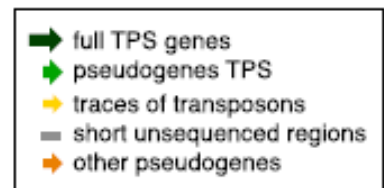
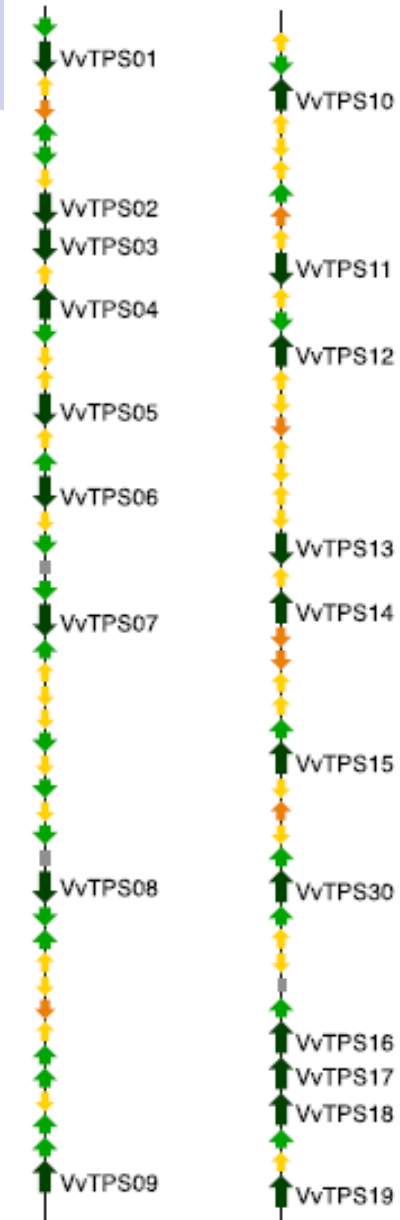
Exemple de la famille des Terpène synthases, impliquées notamment dans l'élaboration des arômes

• 2x plus de gènes de terpène synthase chez la vigne que dans les génomes connus:

- **Vigne : 70** (dont 17 de la sous-famille impliquée dans les arômes floraux)
- **Arabidopsis : 32** (dont 1 de la sous famille impliquée dans les arômes floraux)
- **Peuplier : 50**
- **Riz : 40**

- Famille qui doit évoluer de façon très dynamique chez la vigne
- Caractérisation fonctionnelle de 39 TPS de vigne : substrats et produits très diversifiés

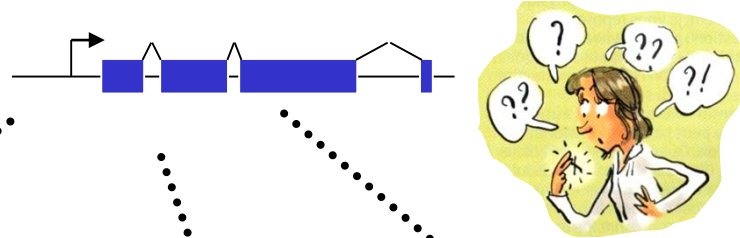
Liens avec la variabilité aromatique des cépages?



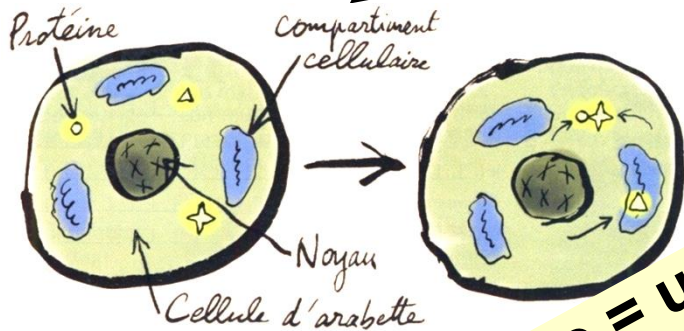
4

Le génomique fonctionnelle :
déterminer la fonction des
gènes et leur réseaux de
régulation

Les outils de la génomique fonctionnelle



Où se trouve la protéine ?
Avec qui interagit-elle ?



Où le gène est-il actif ?
• dans quel tissu ?
• à quel stade ?
• dans quelles conditions ?

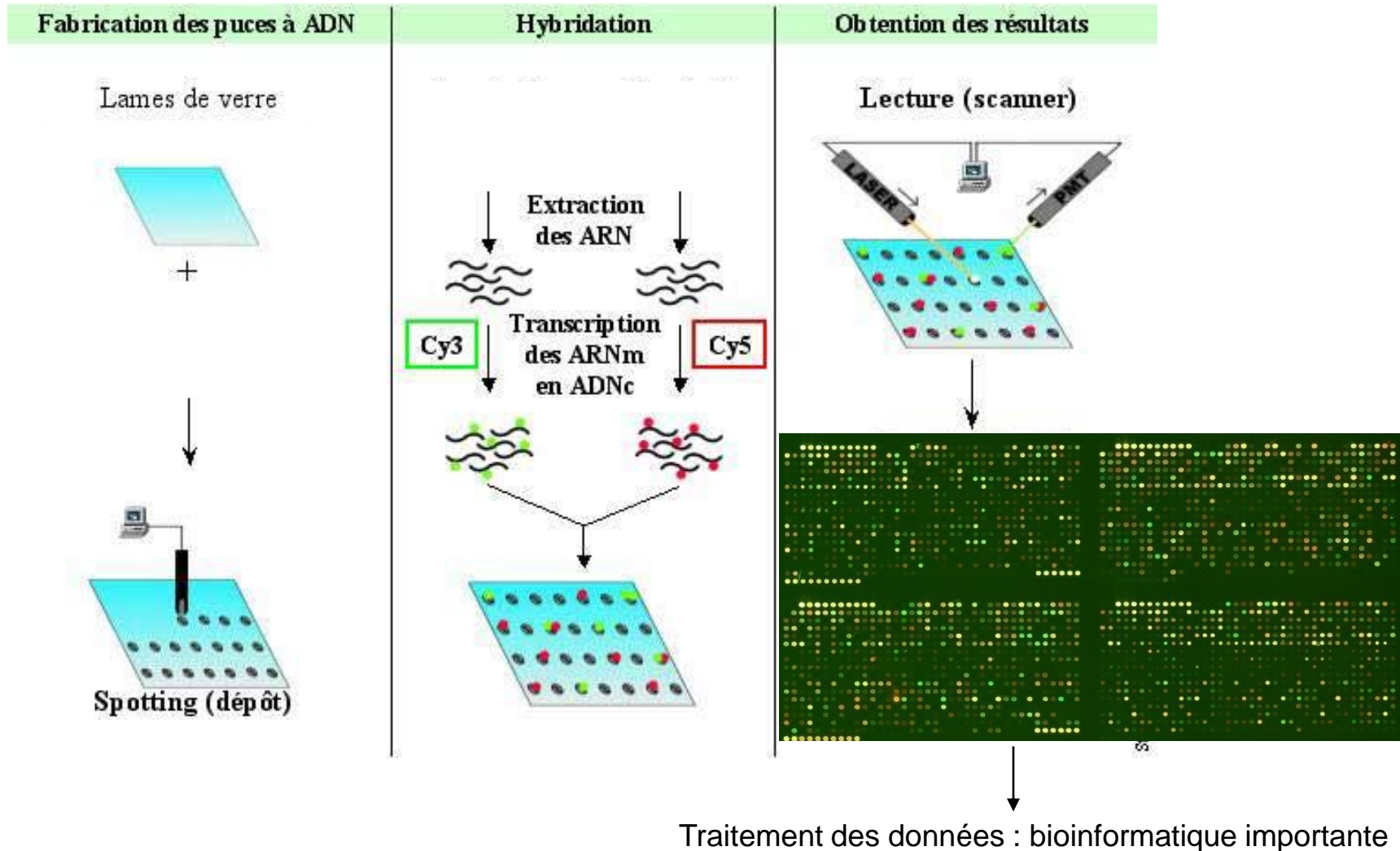
Un gène = une thèse
30000 gènes = 30000 thésards ??!

Que se passe-t-il si
le gène est
manquant ?



Transcriptomique : où les gènes sont-ils exprimés?

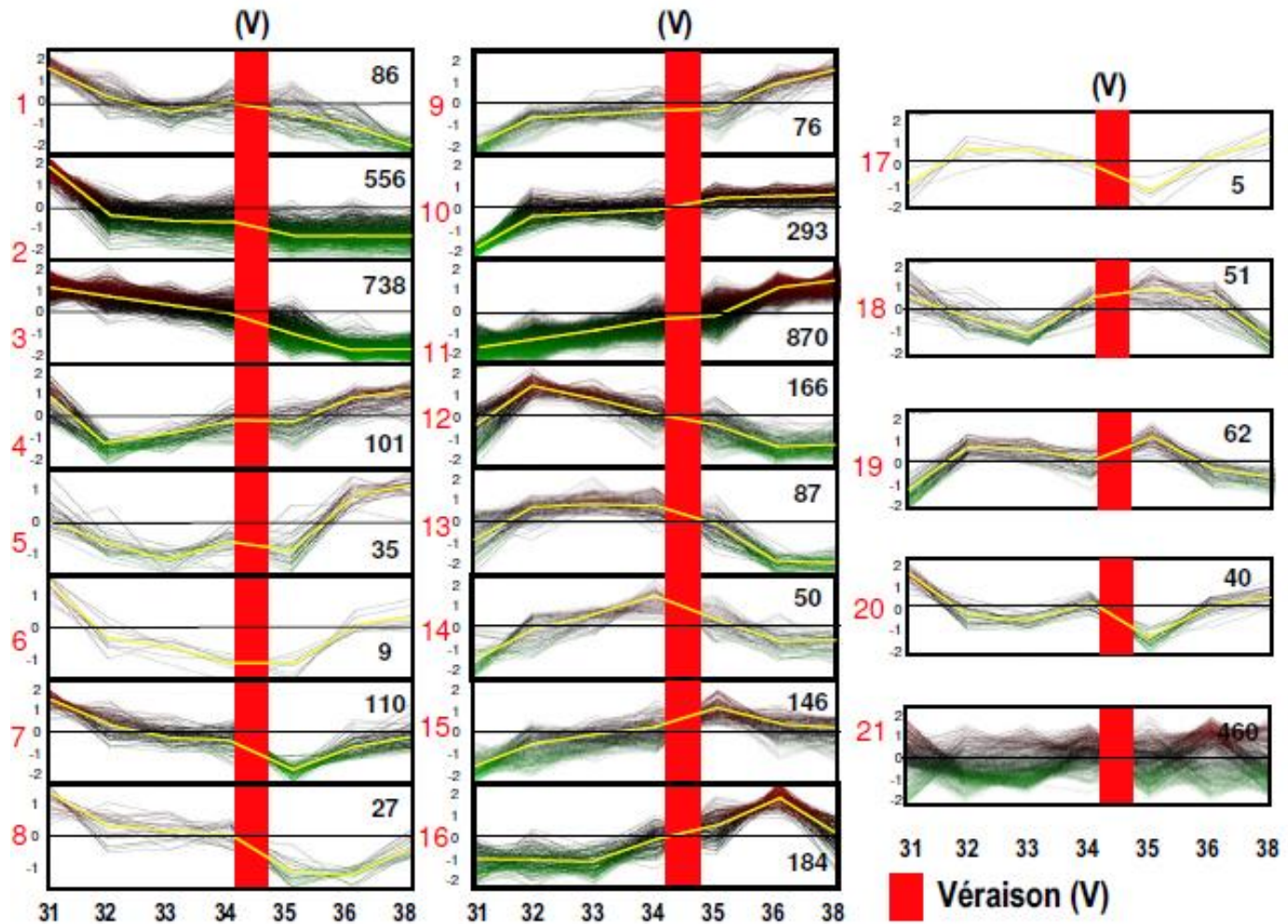
Puces avec des oligonucleotides correspondant à l'ensemble des gènes d'un génome



Traitement des données : bioinformatique importante

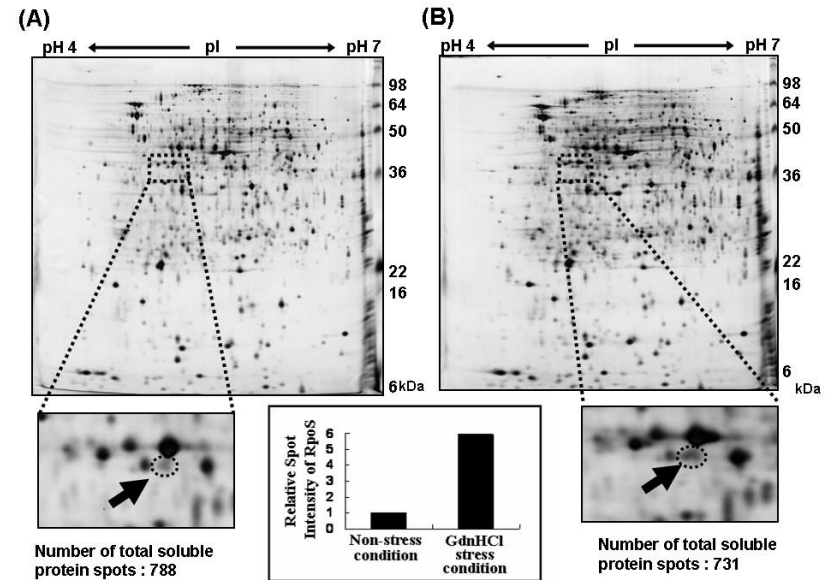
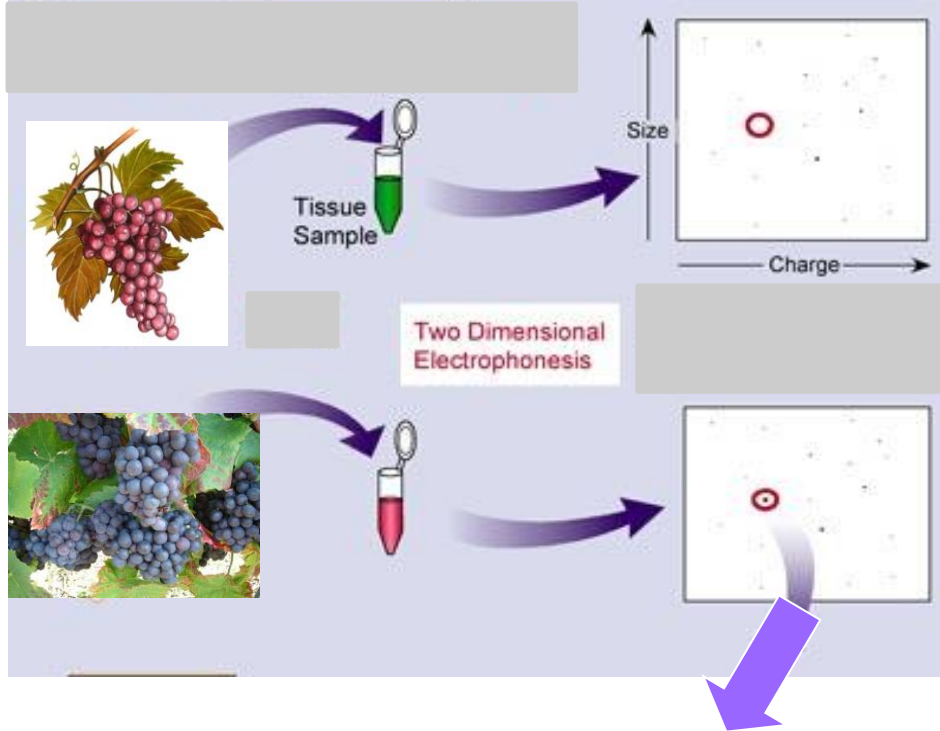
Possibilité de mettre en évidence de réseaux de gènes co-régulés (Deluc et al BMC Genomics 2007)

Au cours de
la maturation
de la baie



Possibilité d'utiliser les nouvelles méthodes de séquençage (Zenoni et al Plant Physiol 2010) : mais encore cher et méthodes statistiques en développement

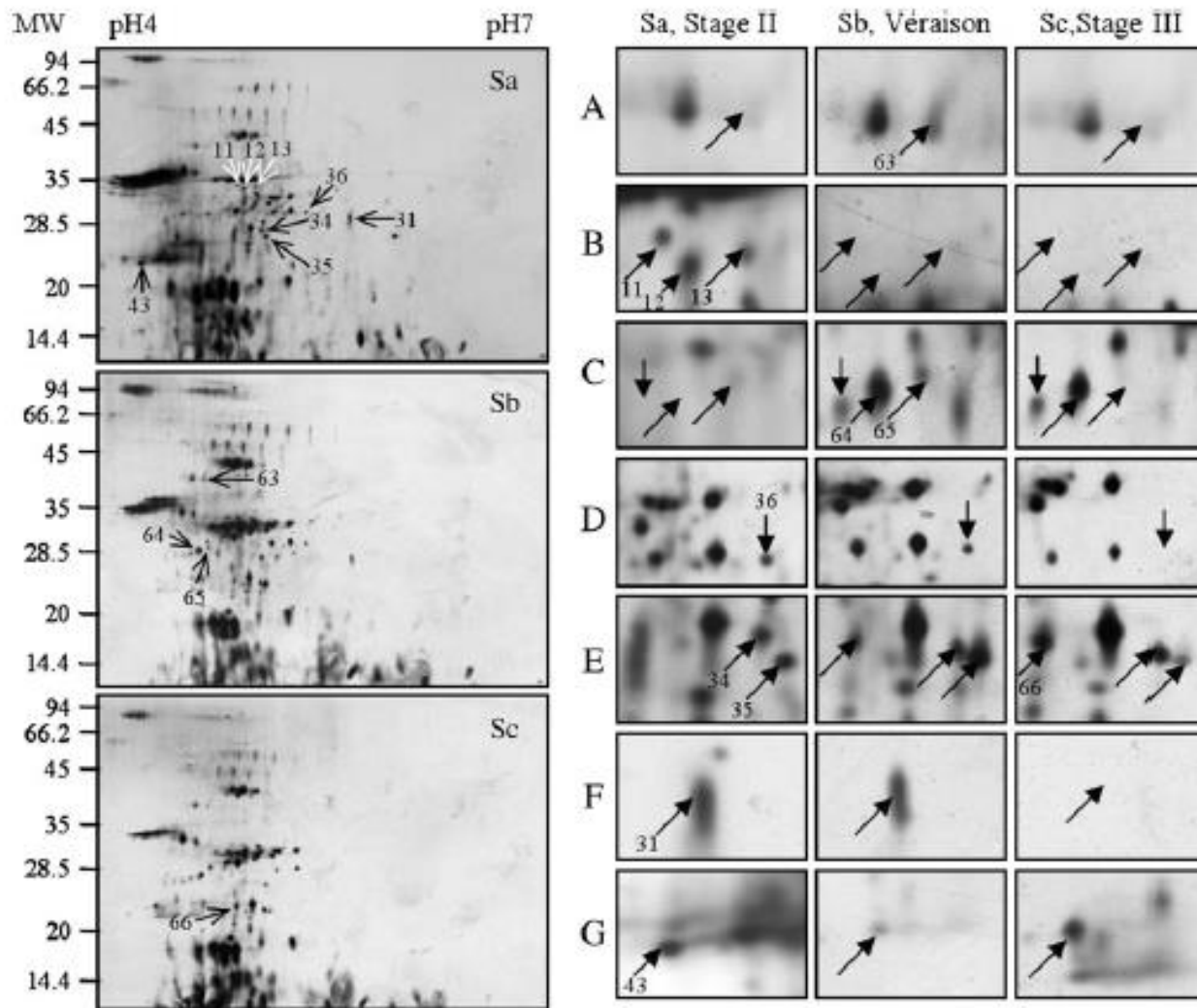
La protéomique: où les protéines sont-elles produites?



Profil peptidiques par Spectrométrie de Masse
Identification par comparaison à des profils connus

- Extraction des protéines différentes suivant leurs propriétés biochimiques
- Pas encore de méthode d'évaluation quantitative
- Plus problématique quand le génome n'a pas été séquencé
- Détection des modifications post-traductionnelles

Etude du protéome des membranes plasmiques des baies à trois stades de développement (Zhang et al, J Exp Bot, 2008)



Identification de 119 protéines en MALDI-TOFF

La protéomique: avec qui les protéines interagissent-elles?

Etude des interactions protéine-protéine

Sensibilité

Couverture

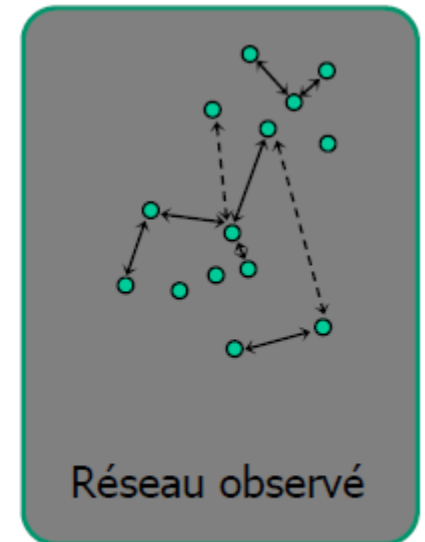
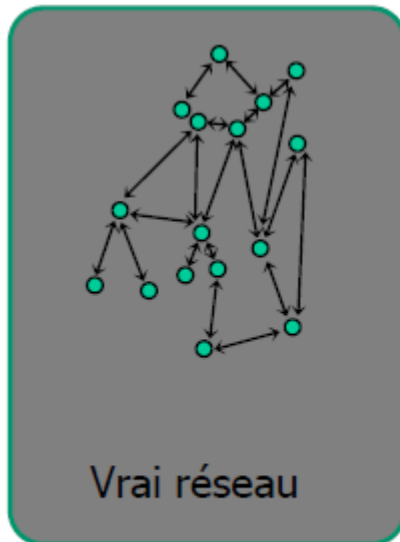
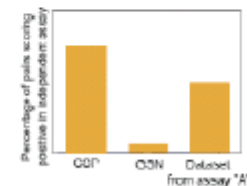


DéTECTABILITÉ

	Interaction 01	Interaction 02	Interaction 03	Interaction 04	Interaction 05	Interaction 06	Interaction 07	Interaction 08
Assay 1	■	■	■	■	■	■	■	■
Assay 2	■	■	■	■	■	■	■	■
Assay 3	■	■	■	■	■	■	■	■

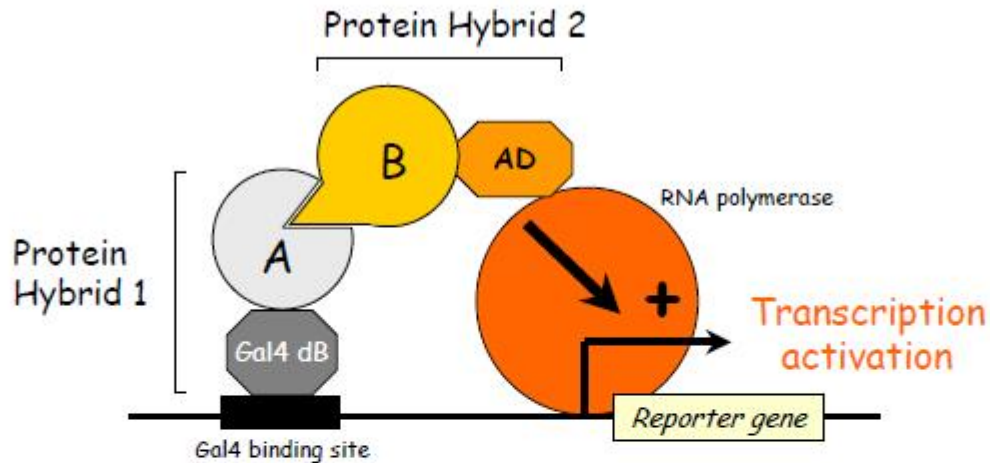
Multiple assays

Spécificité

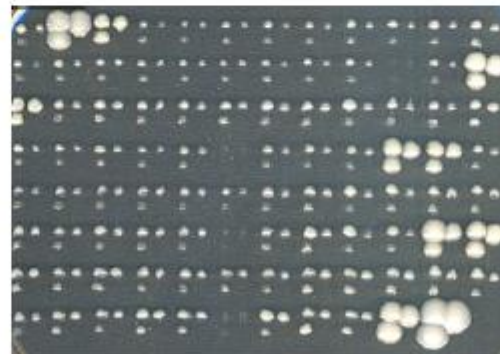


La protéomique: avec qui les protéines interagissent-elles?

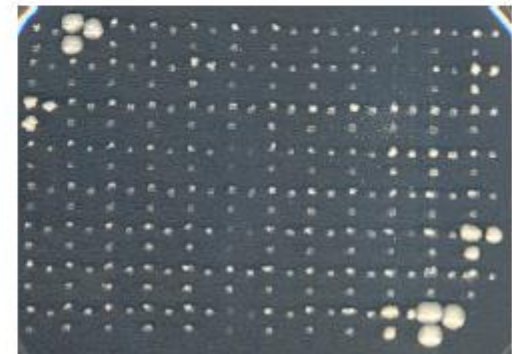
Double hybride chez la levure



Xgal



His-



Ura-

30 000 gènes => $9 \cdot 10^8$ interactions deux à deux à tester!

La génétique inverse : quel est l'effet sur la plante lorsqu'un gène n'est plus fonctionnel?

Mutagenèse chimique/physique: U.V, rayon X, EMS ...
substitutions, délétions;
mutations ponctuelles difficiles à localiser dans le génome

Mutagenèse insertionnelle : T-DNA (ou Transposon)
mutants perte de fonction;
Mutants étiquetés
souvent plusieurs T-DNA/transposons dans un transformant

Chez la vigne

✓ Essais de collection
mutants EMS sans
succès

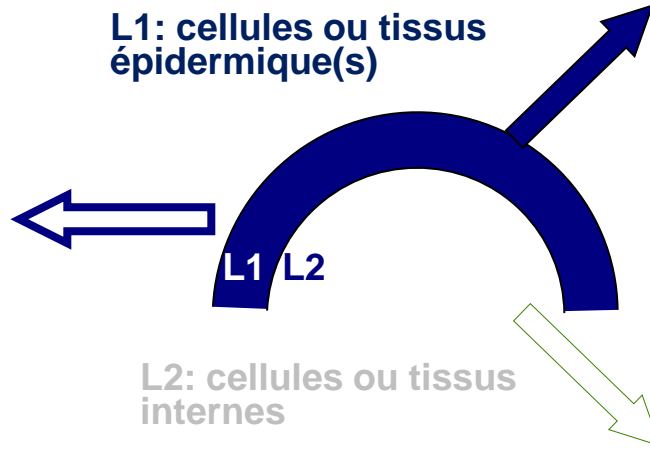
✓ Problème d'efficacité
de transformation
✓ Collection difficile
à maintenir

Il existe des mutants naturels chez la vigne, mais ce sont souvent des chimères



Pinot gris

L1: cellules ou tissus épidermique(s)



L2: cellules ou tissus internes



Plante obtenue par
régénération d'embryons
somatiques L1
Type Pinot noir



Plante issue
d'autofécondation
(L2)
Type Pinot blanc

5

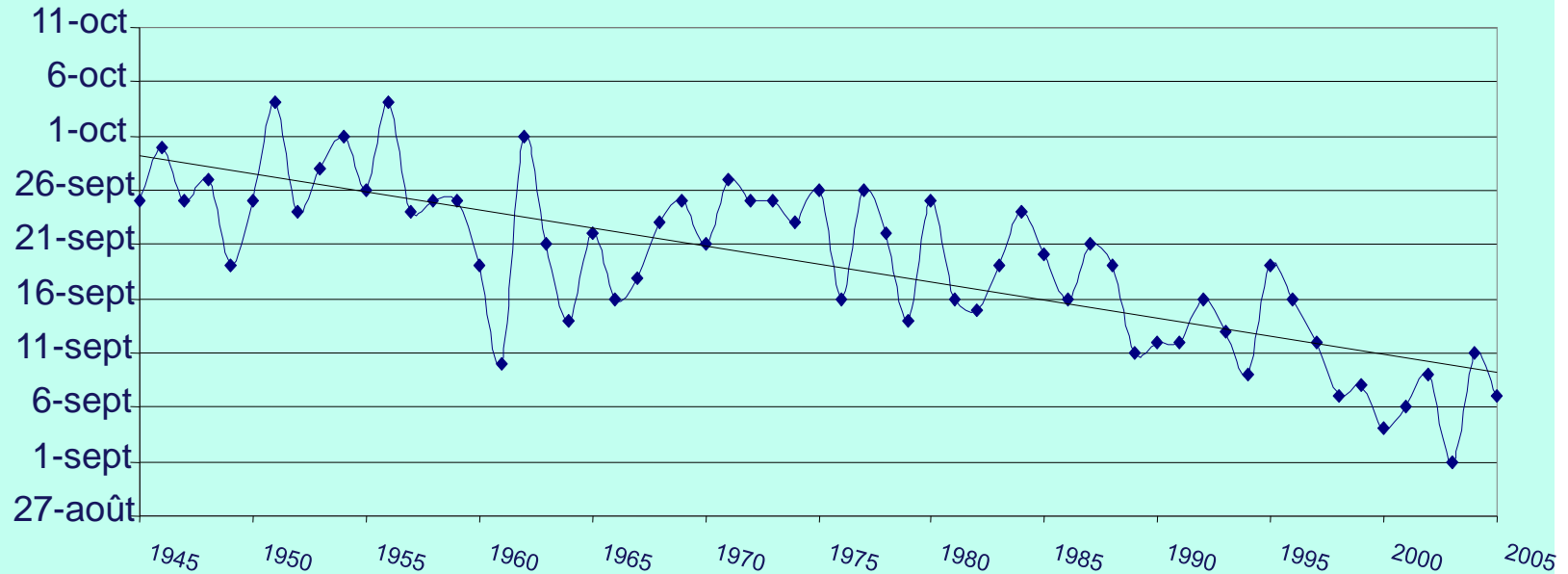
Champs d'Applications

Deux problèmes essentiels

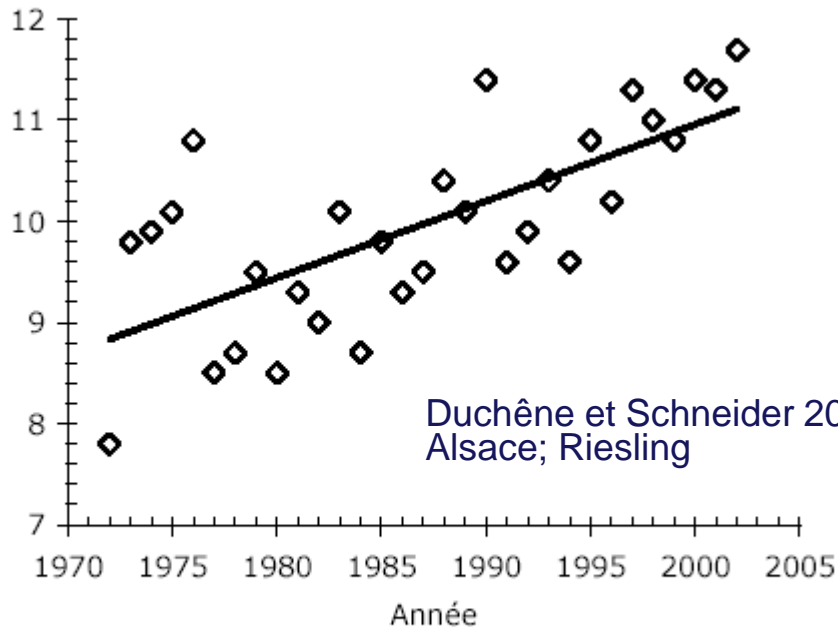
- Qualité en relation avec l'adaptation à l'environnement
- Contrôle des maladies

DATE DE DEBUT DES VENDANGES A CHATEAUNEUF DU PAPE depuis 1945

Source : Service technique Inter Rhône

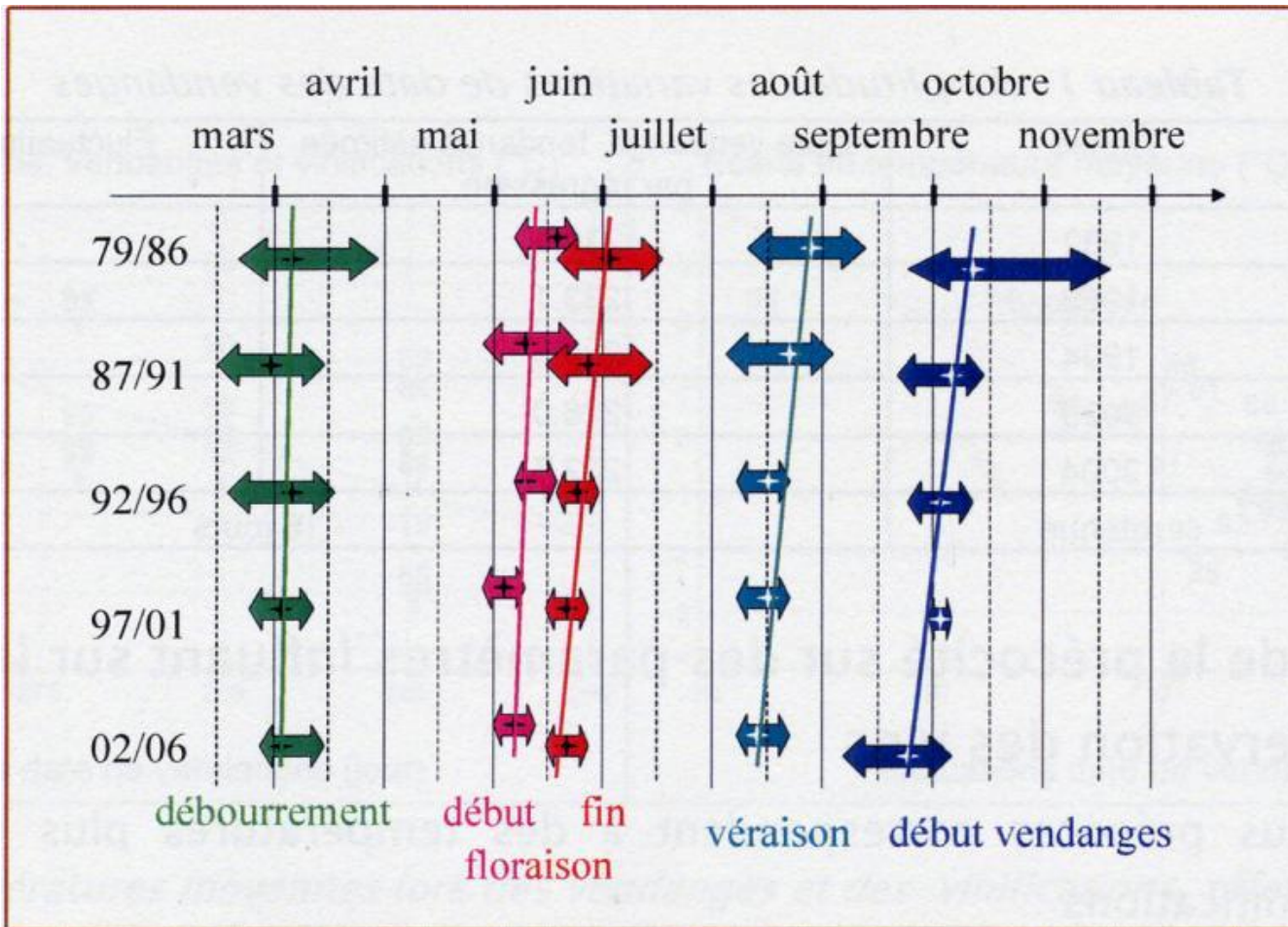


Degré alcoolique probable
à la récolte (%vol)



Synchronisation entre
maturité phénolique et
maturité alcoolique

J-P Gotouly, ISVV, 2008



- Raccourcissement du cycle végétatif
- Précocité des stades phénologiques
- Décalage du cycle vers les périodes chaudes de l'été

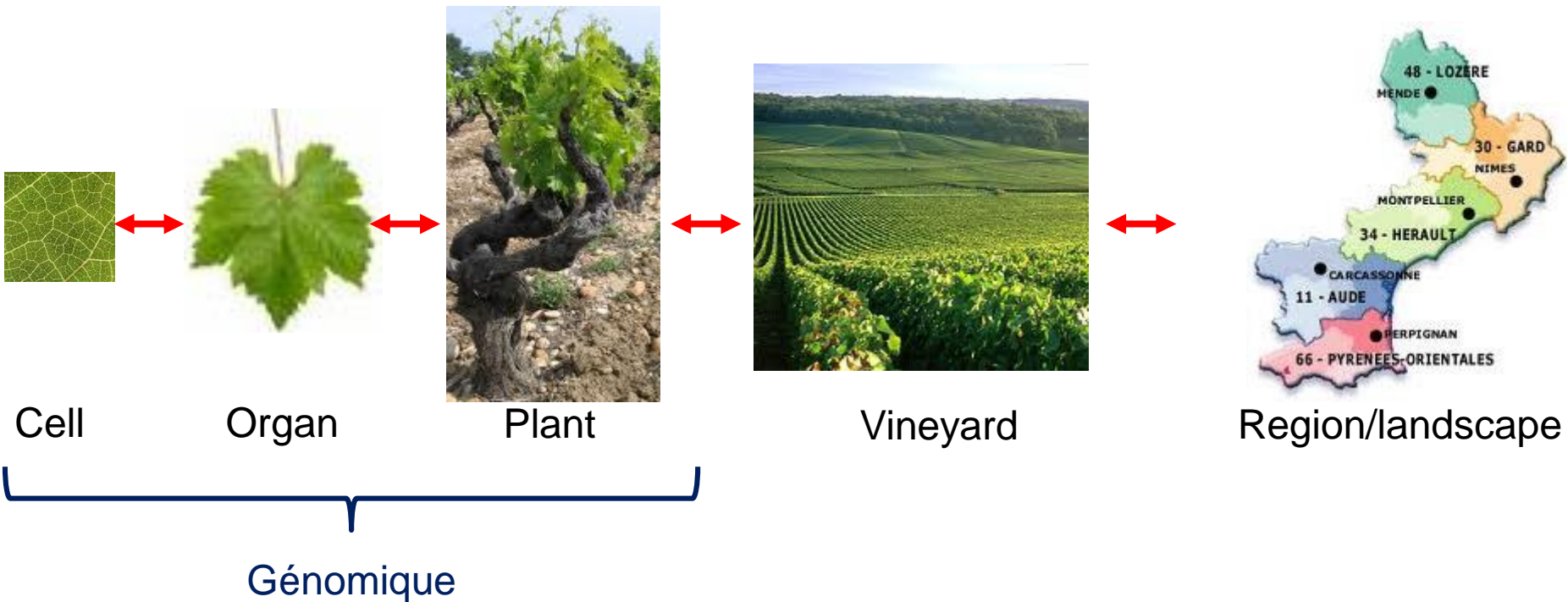
Biologie intégrative

Comment prédire le phénotype à partir des informations disponibles : génotype, environnement ? Comment construire de nouveaux idéotypes variétaux ?



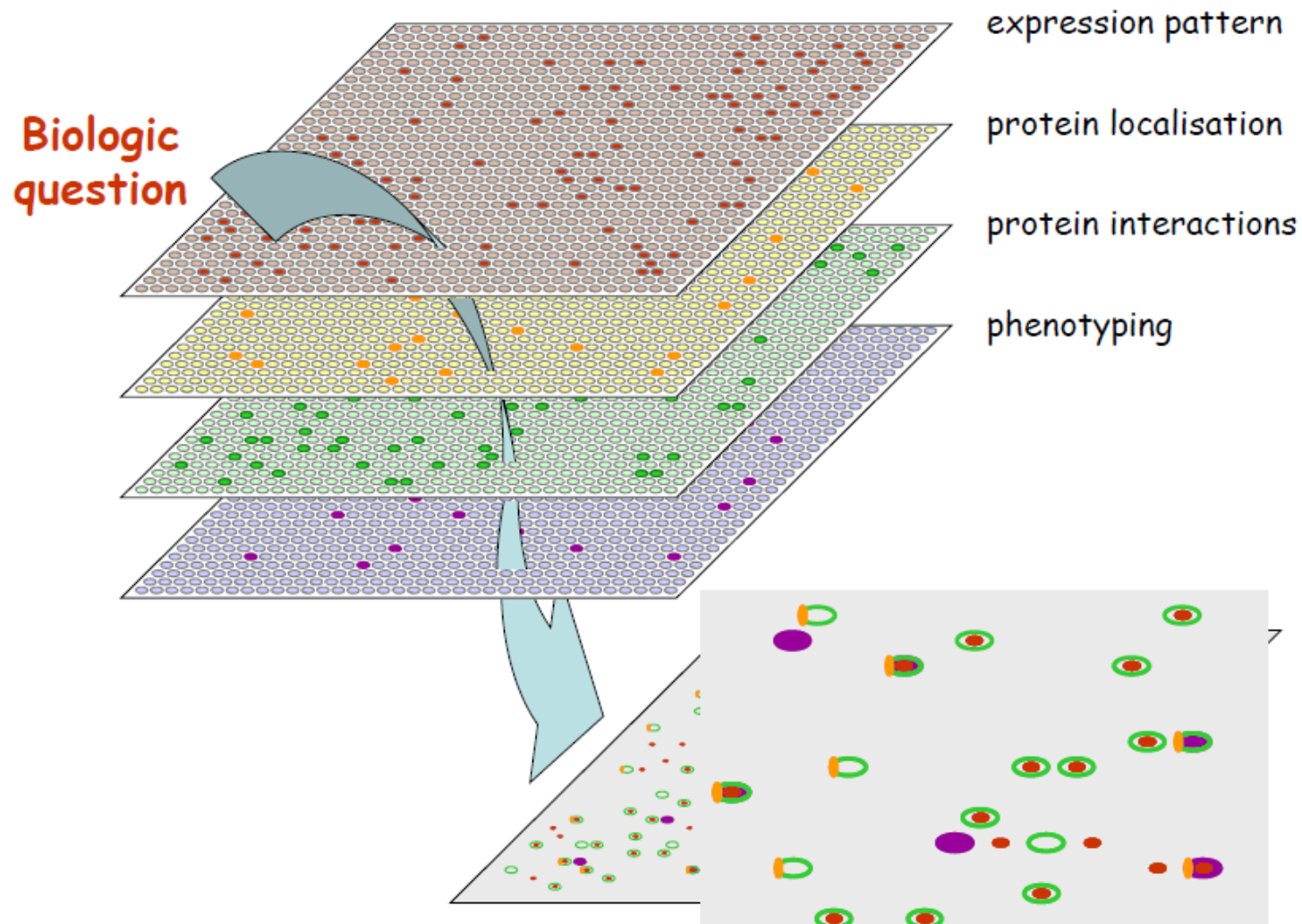
Modélisation à différentes échelles

Tardieu & Tuberosa 2010
Marguerit et al 2012



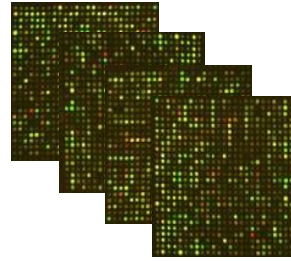
From S. Delrot

Intégration de données: vers la génomique prédictive



Intégration de données: vers la génomique prédictive

Une ressource
transcriptome
(Ex: 8000 hybridations chez
A. thaliana dans CatDB)



RNA-seq

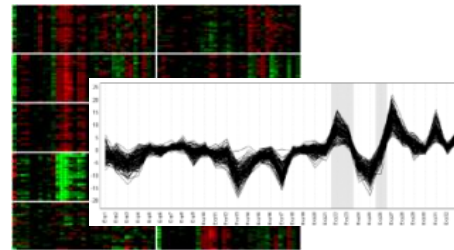


Modèles statistiques de mélange

interactome



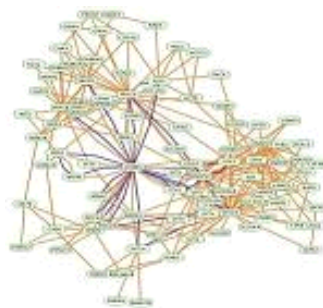
phosphoprotéome



Clusters de gènes co-régulés
dans des conditions données

*Modèle
Graphique Gaussien*

Réseaux fonctionnels



Informatique, mathématiques et bases de données

Intégration de données de transcription de gènes, de protéome et de métabolome au cours de la sur-maturation de la baie (Zamboni et al Plant Physiol 2010)

Regulation of transcription

- ◆ MYB
- ◆ bHLH
- ◆ repress protein

transcrits

Withering
Stress responses

protéines

Oxidative stress

- ◆ catalase
- ◆ glutathione-S-transferase
- ◆ glutaredoxin
- ◆ ascorbate peroxidase
- ▲ thionin-like protein

Water stress

- ◆ dehydration-induced protein
- ◆ trehalase synthase
- ▲ osmotin-like protein

Response to biotic stimulus

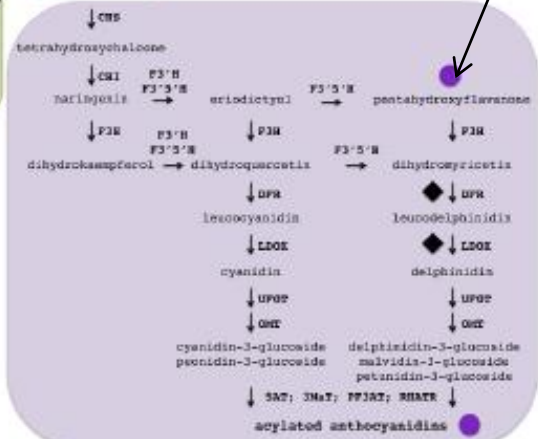
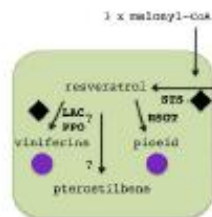
- ◆ Avr9/CF-9 rapidly elicited proteins
- ◆ disease resistance protein
- ◆ pathogen-related protein

Phenolic secondary metabolites involved in defence responses

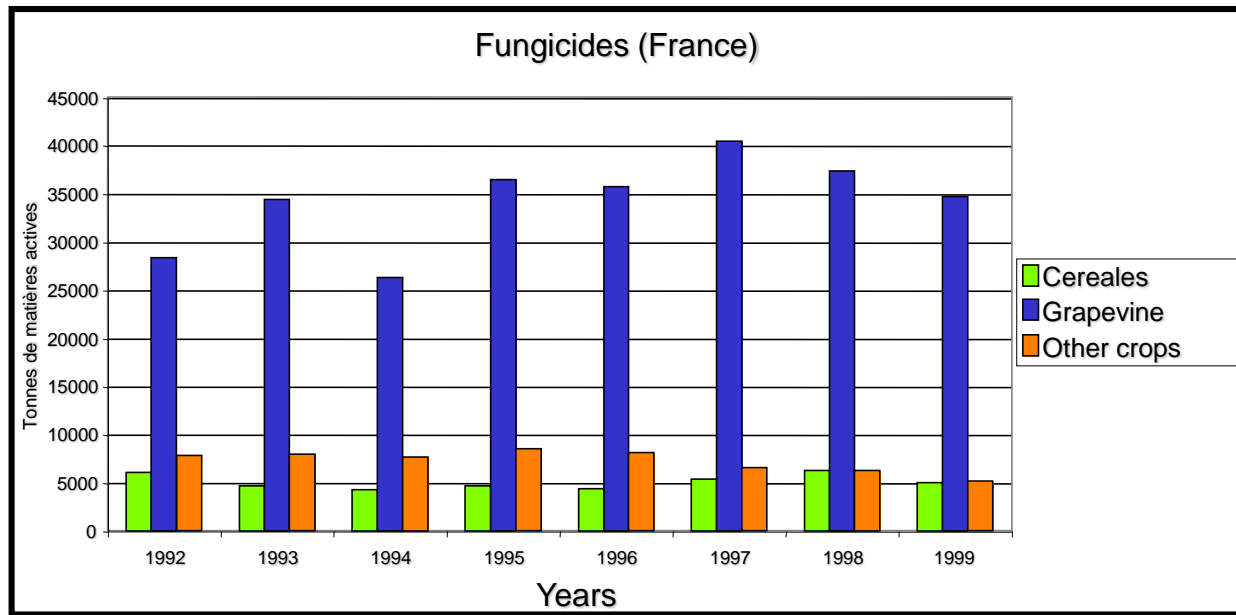
shikimate pathway

- ◆ DAHP; DSM; EPS
- ◆ phenylalanine
- ◆ PAL
- ◆ cinnamic acid
- ◆ C4H
- ◆ 4-coumaric acid
- ◆ 4CL
- ◆ 4-coumaroyl-CoA

métabolites



Intérêt de développer des variétés résistantes aux maladies fongiques



Utilisation de
fungicides en
agriculture

Ces résistances doivent être durables

⇒ combiner plusieurs mécanismes de défense/résistance dans une même variété

⇒ Combiner les approches génétiques, agronomiques, phytosanitaires

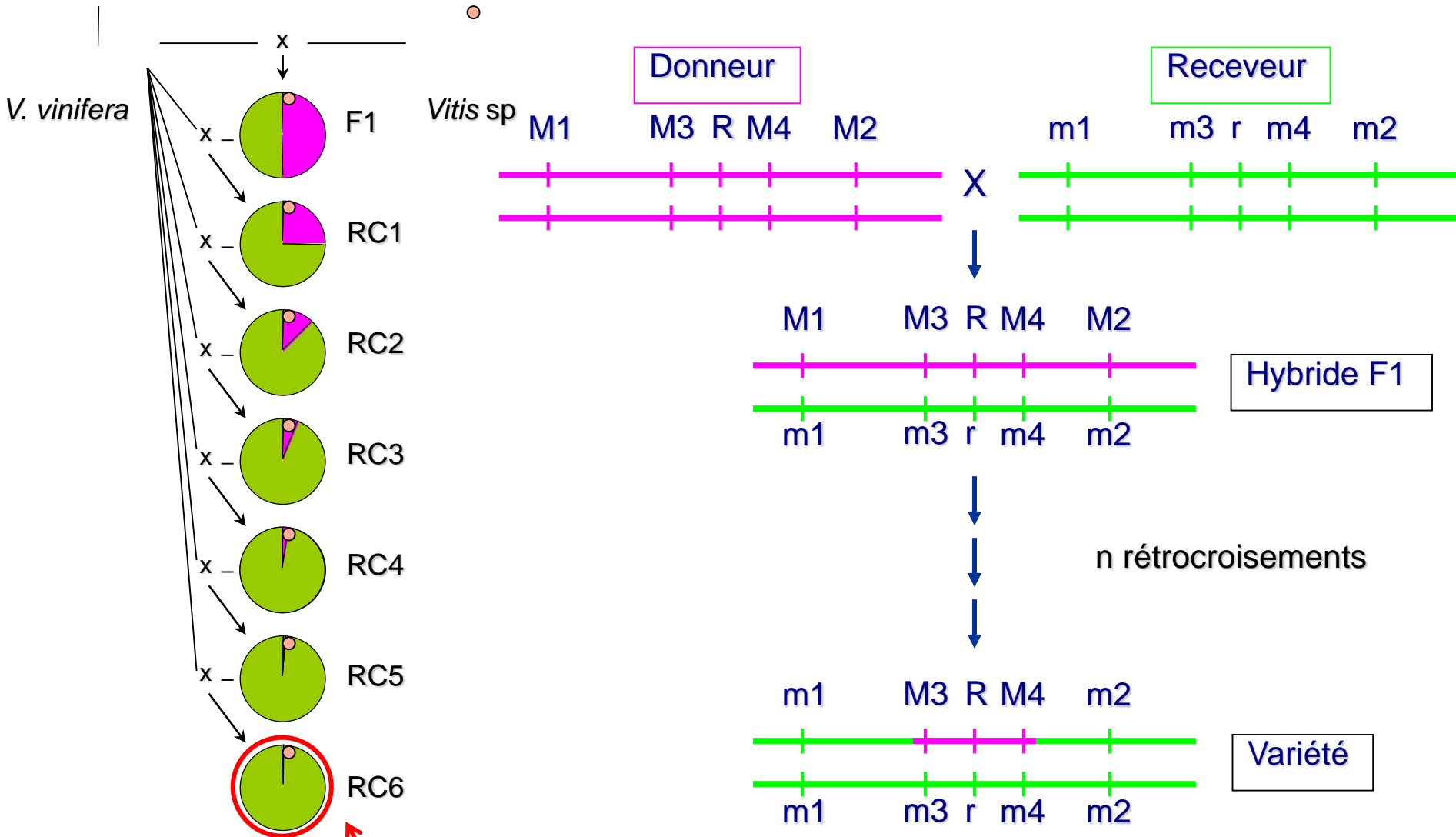
Les sources de résistance sont dans le complexe des Vitaceae

Pathogen	Source of resistance	Reference
<i>P. viticola</i>	<i>V. riparia</i>	Marguerit <i>et al</i> 2009
		Grando <i>et al</i> 2003
	Complex hybrid	Fischer <i>et al</i> 2004
<i>E. necator</i>	<i>M. rotundifolia</i>	Barker <i>et al</i> 2005
	<i>M. rotundifolia</i>	Eibach <i>et al</i> 2007
	<i>V. Vinifera</i>	Hoffmann <i>et al</i> 2008; Coleman <i>et al</i> 2009
<i>X. fastidiosa</i>	<i>V. arizonica</i> '	Krivanek <i>et al</i> 2006
	<i>V. arizonica</i> × <i>V. candidans</i>	Riaz <i>et al</i> 2008
<i>D. vitifolia</i>	<i>V. riparia</i> × <i>V. cinerea</i>	Zhang <i>et al</i> 2009

Travail avec des croisements interspécifiques : problèmes?

Aller de la sélection assistée par
marqueurs à la sélection génomique

Sélection assistée par marqueurs de nouvelles variétés durablement résistantes aux maladies

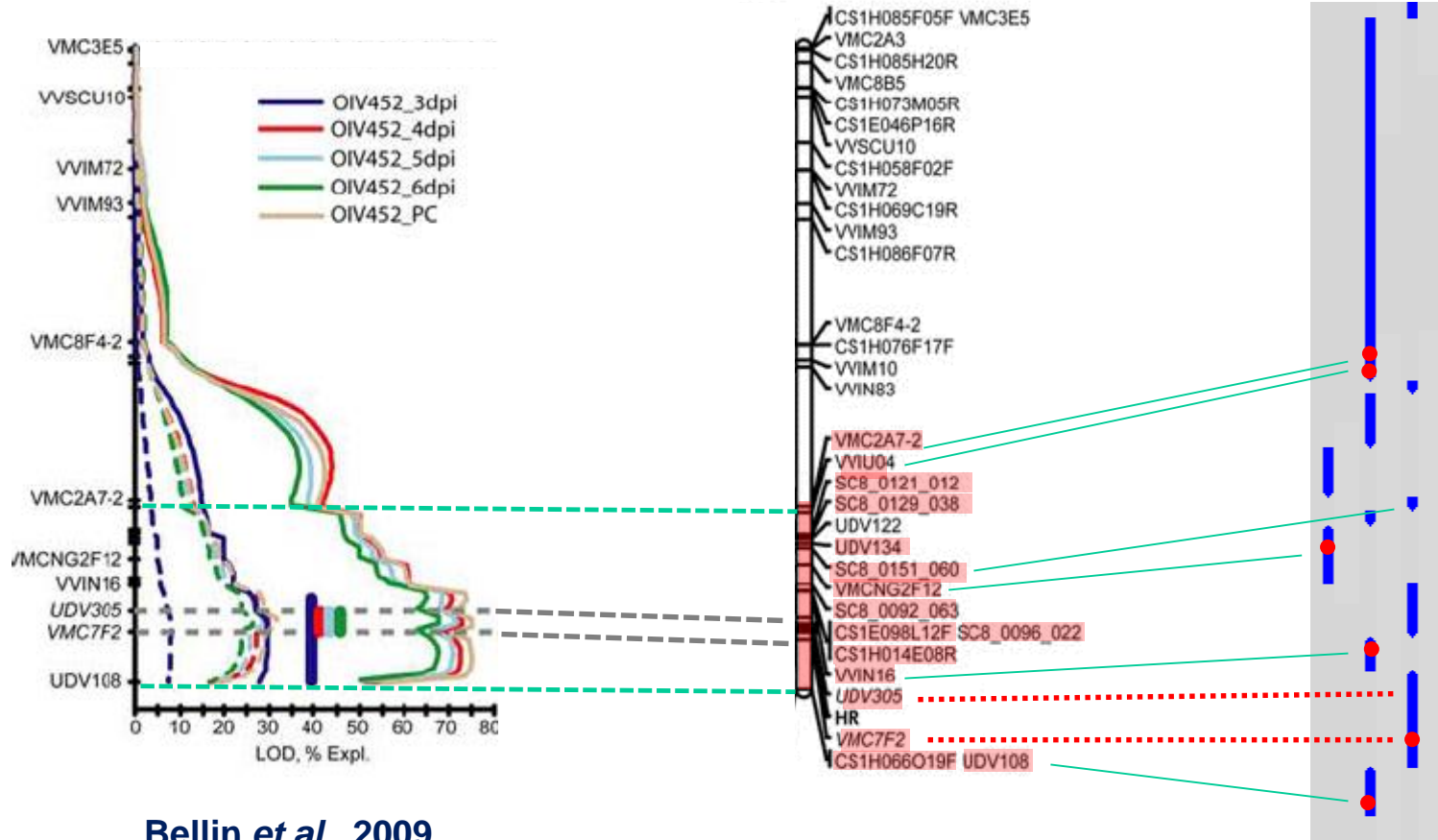


Génomique comparative entre différentes espèces de la famille des Vitaceae

QTL
Chromosom
18

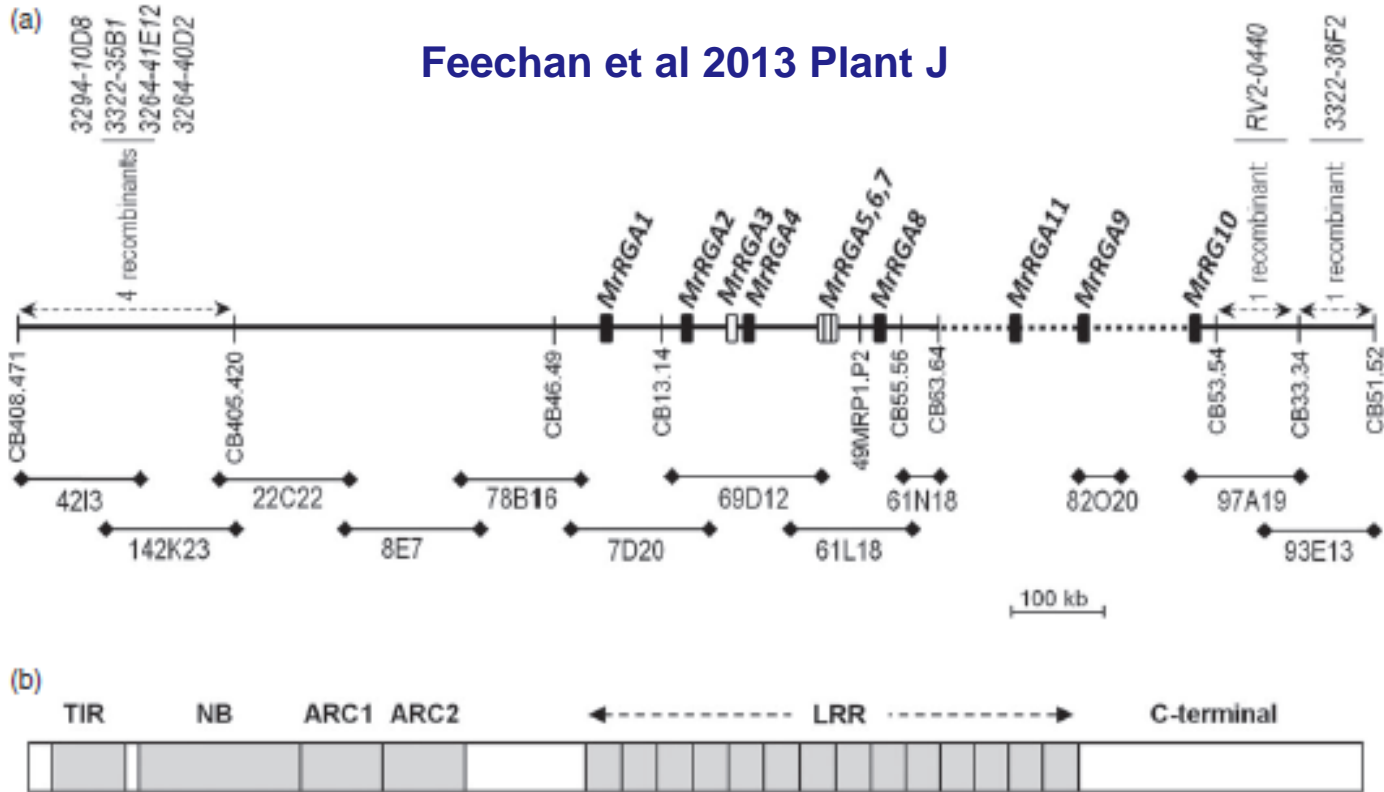
Reference
genetic map

Genome
sequence



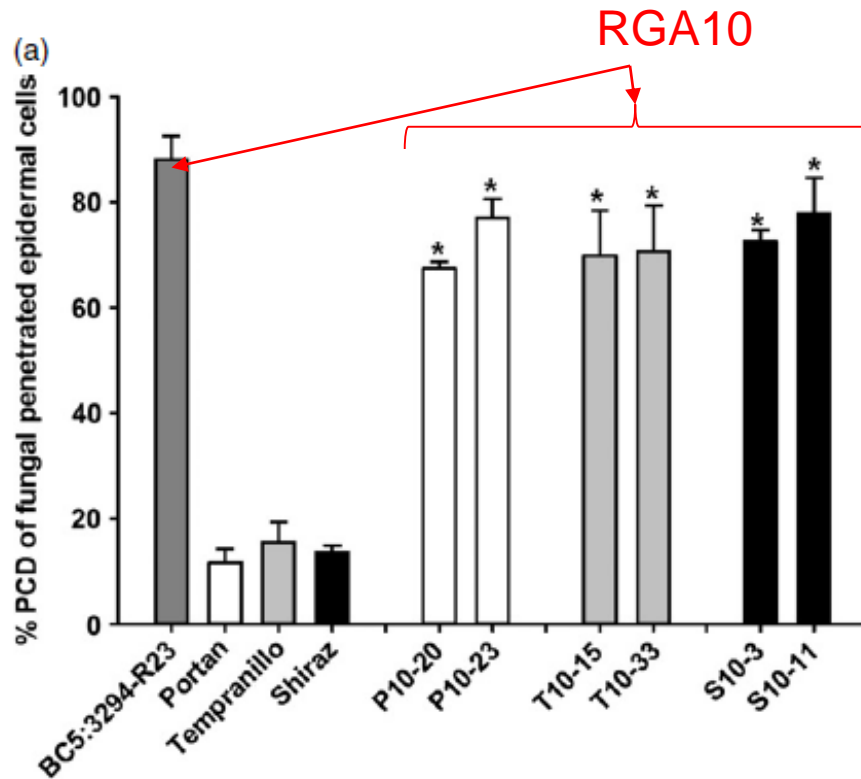
Bellin *et al.*, 2009
Downy mildew resistance from cv. Bianca

Caractérisation moléculaire d'une source de résistance à l'oïdium et au mildiou (16 ans de collaboration INRA-CSIRO)



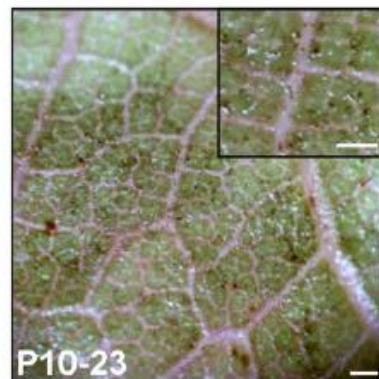
- Résistance totale à l'oïdium (Run1) et partielle au mildiou (RpV1) provenant de *Muscadinia rotundifolia* cartographié génétiquement grâce à des populations générées à l'INRA par Alain Bouquet (plus d'un milliers de plantes criblées)
- Une banque de grands fragments d'ADN de muscadine alignée sur la carte génétique et séquencée (CSIRO) => mise ne évidence d'un groupe de gènes candidats

Isolement du gène de résistance à l'oïdium: RGA10 = Run1

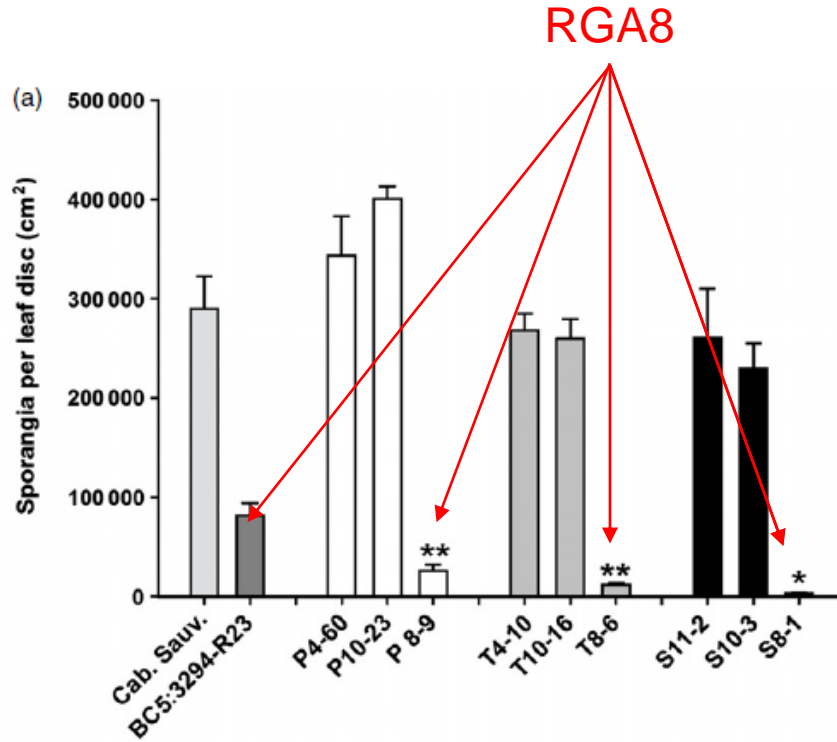


Des plantes sensibles à l'oïdium (e.g. cv Portan) dans lesquelles on a transféré le gène RGA10 provenant de la Muscadine deviennent totalement résistantes à l'oïdium

% de cellules pénétrées par une conidie qui expriment une mort cellulaire programmée similaire avec le témoin résistant (BC5-3294-R23)

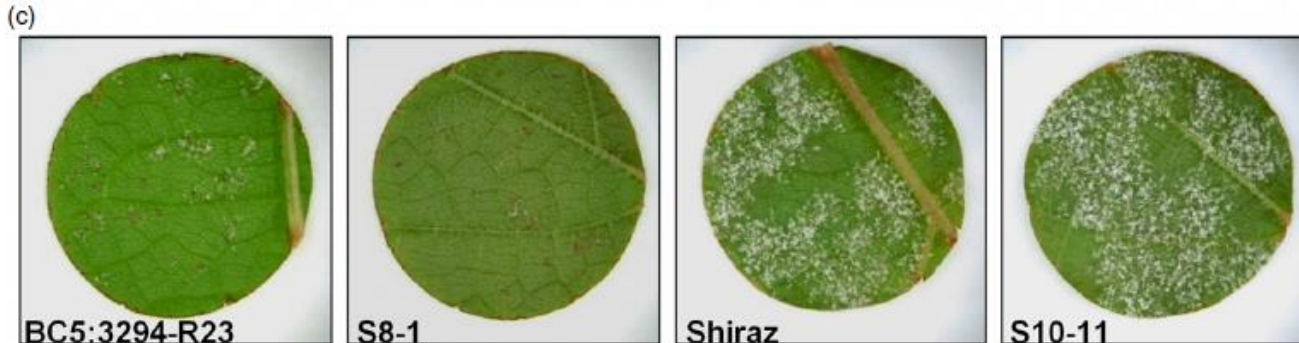


Isolement du gène de résistance au mildiou: RGA8 = Rpv1



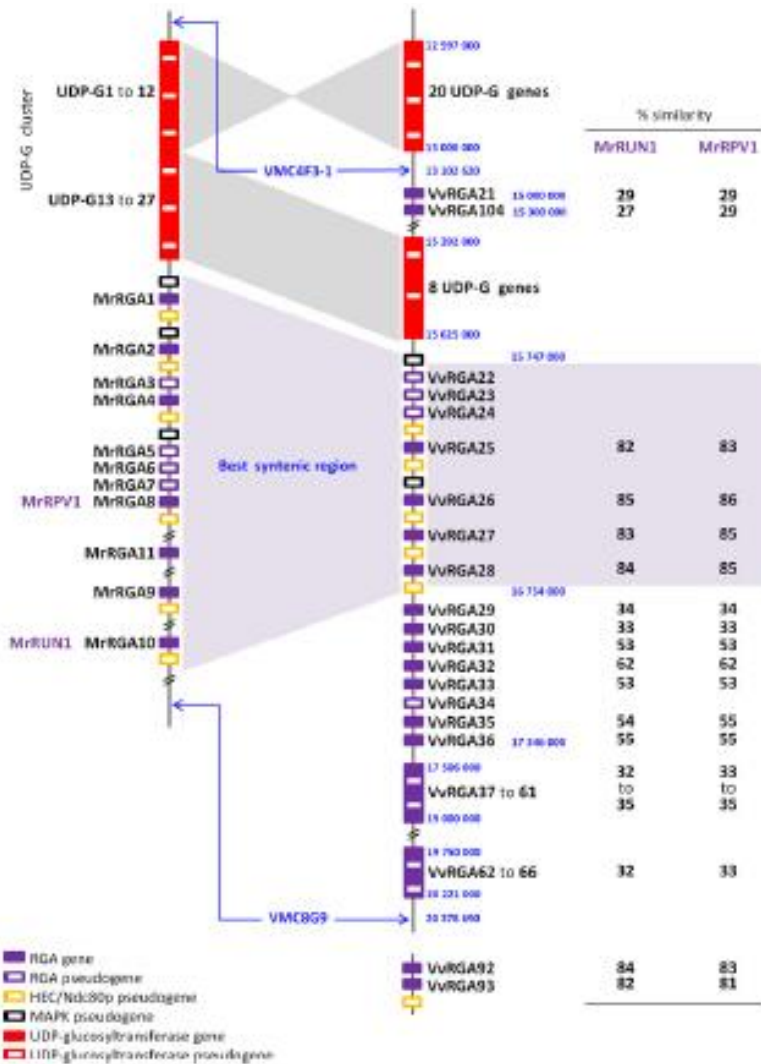
Des plantes sensibles à l'oïdium (e.g. cv Portan) dans lesquelles on a transféré le gène RGA8 provenant de la Muscadine présentent de très hauts niveaux de résistance au mildiou

% sporanges similaires voir même plus bas que chez le témoin résistant (BC5-3294-R23)



Les regions homologues chez *V. vinifera* et *M. rotundifolia* sont très différentes

Muscadinia *V. vinifera*



Une inversion et une évolution différente du cluster de gènes NBS-LRR auquel appartiennent Run1 et Rpv1

Création variétale

Caractères pour lesquels des études génétiques ont été menées

Résistance (Oïdium, Mildiou)

Sexe

Fertilité

Arome Muscat

Taille de la baie

La couleur

La composition en tannins et anthocyanes

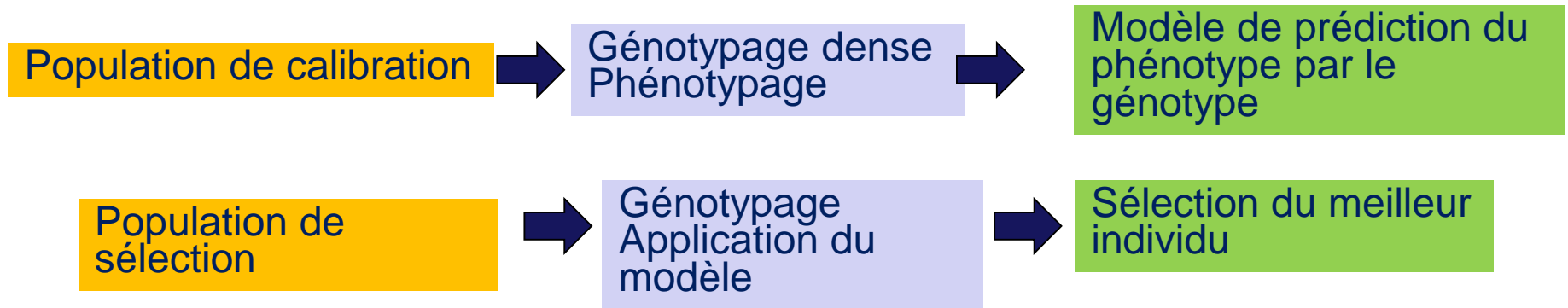
L'apyrénie

L'acidité, la compacité, la longueur et la largeur de la grappe.....

- Résistance aux maladies : travail dans un complexe d'espèces => génomique comparative
- Cumul dans une même variété de nombreux critères à déterminisme polygénique => nombreux marqueurs nécessaires et demande du temps!
- Durabilité de la résistance dans le temps => cumul de plusieurs gènes de résistance à un même pathogène nécessaire
- Nouvelles approches à tester pour le choix des marqueurs et le matériel biologique (élargissement des bases génétiques, cycles courts...)

Développement de la sélection génomique chez la vigne?

Concept développé chez les animaux et qui a complètement révolutionné l'organisation de la sélection



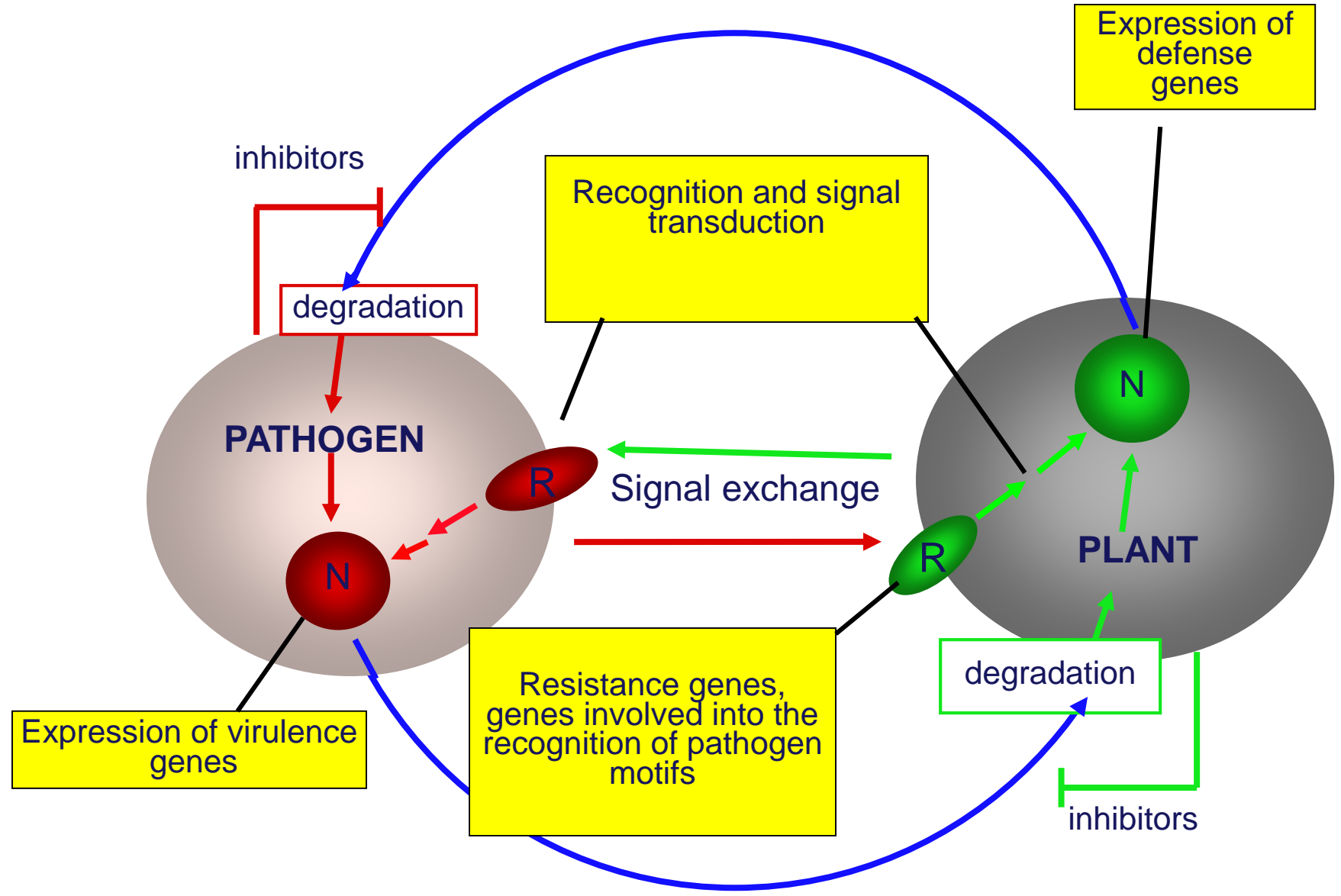
Meuwissen et al, Genetics, 2001

Cela reste à tester chez les plantes en fonction de la structure de leurs populations, le fait qu'on travaille souvent en croisements interspécifiques...

Développement d'outils qui permettent d'analyser en une fois plusieurs dizaines ou centaines de milliers de marqueurs => re-séquençage de variétés/espèces diverses avec les nouvelles générations de séquençage

Etude de l'interaction entre deux organismes : interactions vigne pathogènes (ou auxiliaires!)

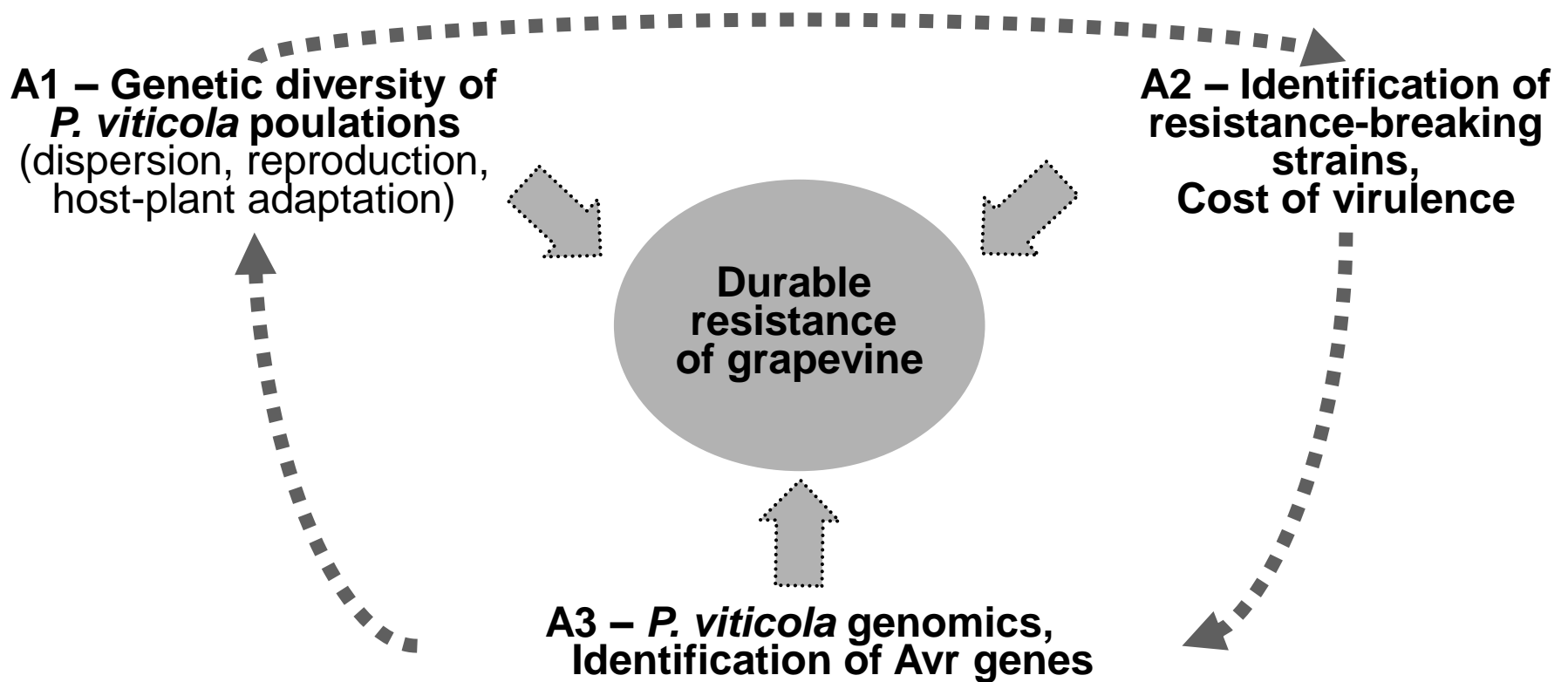
L'interaction entre une plante et un agent pathogène



Pour certaines maladies, il n'y a pas de source de résistance dans les ressources génétiques vigne

Il est important de développer des ressources génomiques pour l'étude des pathogènes :

- Identification des gènes clefs dans leur cycle infectieux

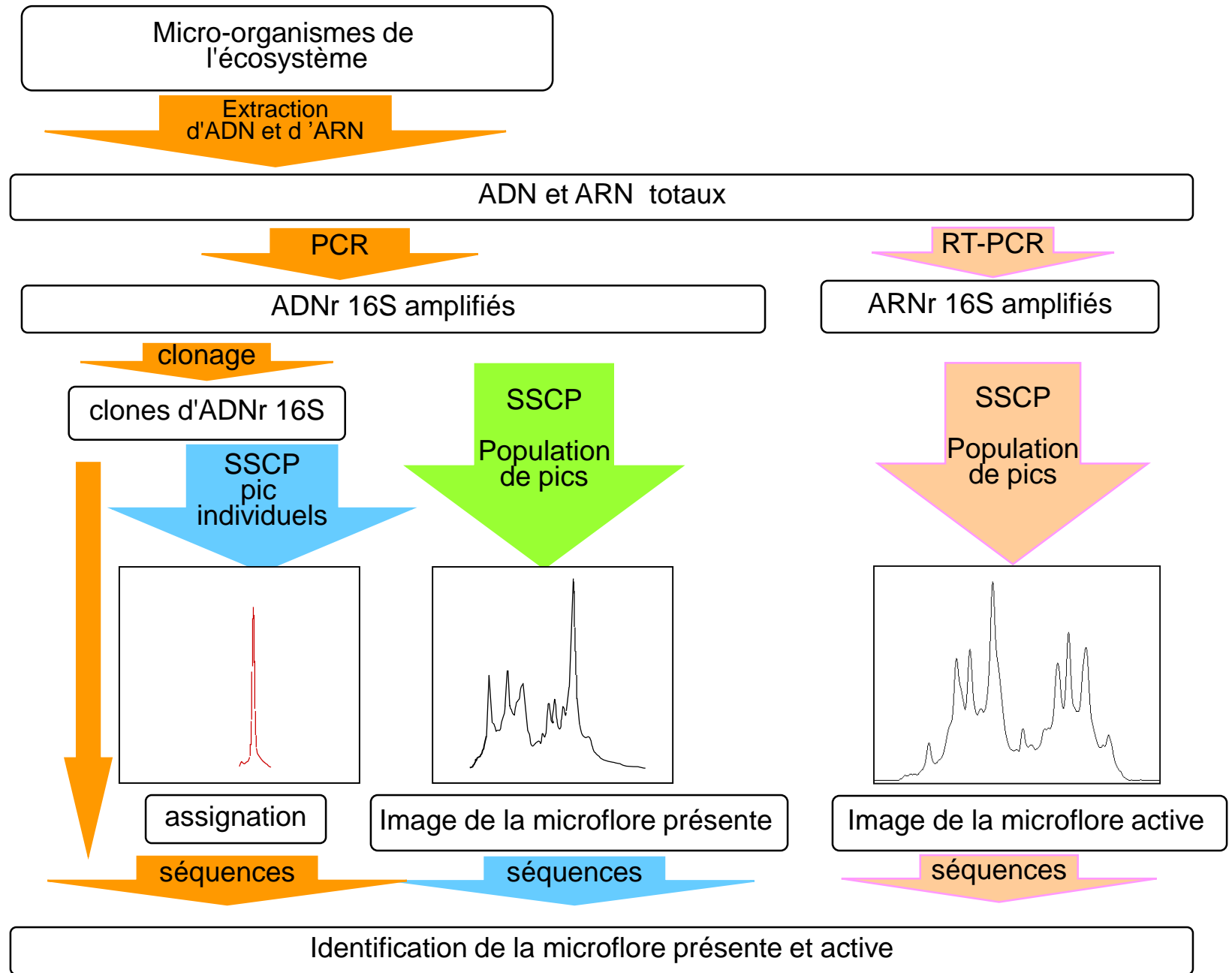


Comparative genomics of pathogens

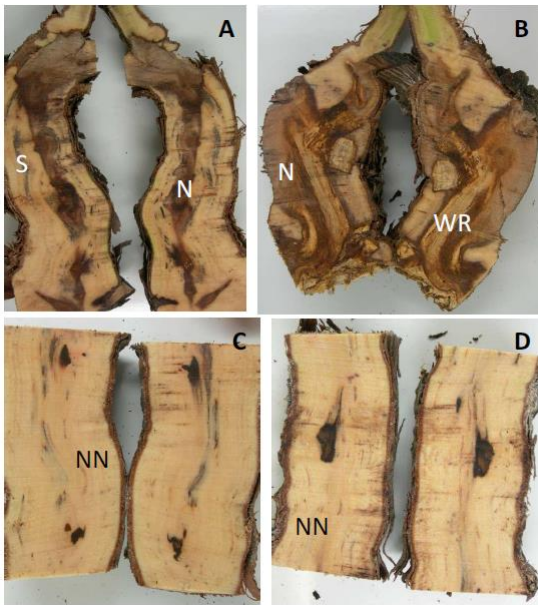
Pour certaines maladies, il n'y a pas de source de résistance dans les ressources génétiques vigne

Il est important de développer des ressources génomiques pour l'étude des pathogènes :

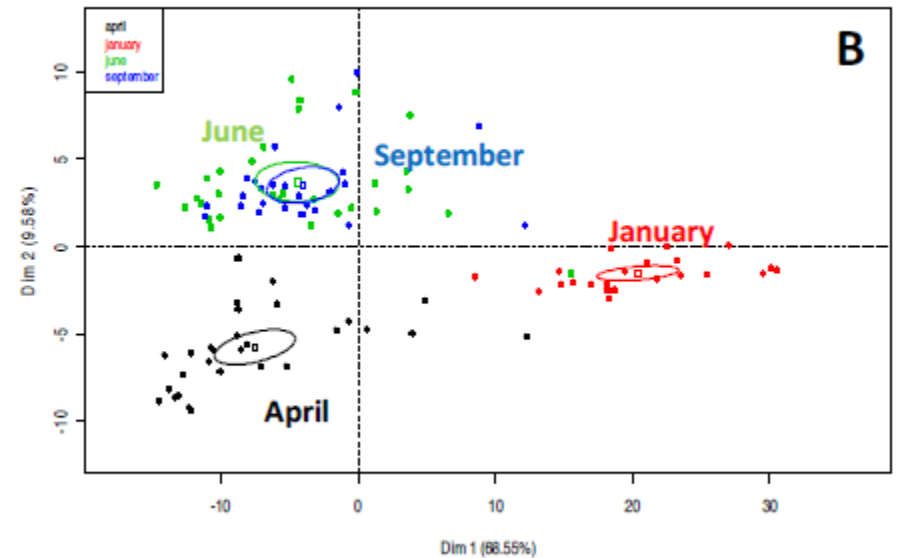
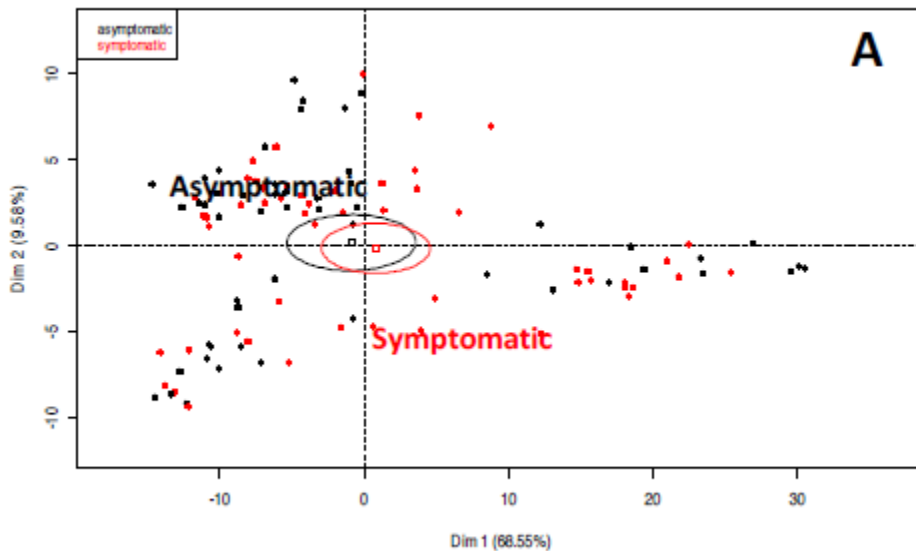
- Identification des gènes clefs dans leur cycle infectieux
- Compréhension des syndromes complexes (approches métagénomiques)



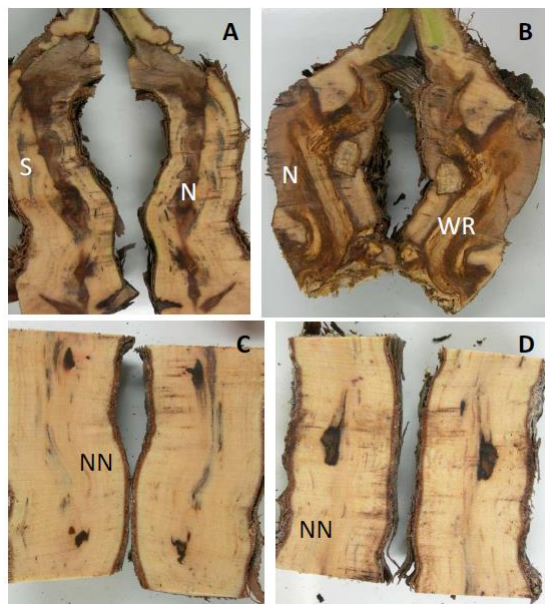
Asymptomatique Symptomatique



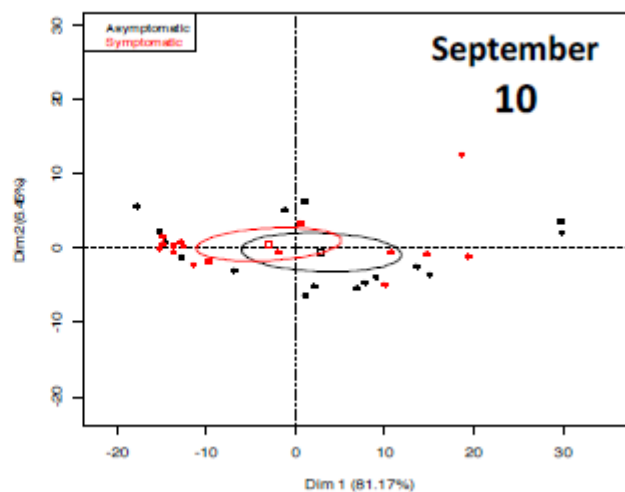
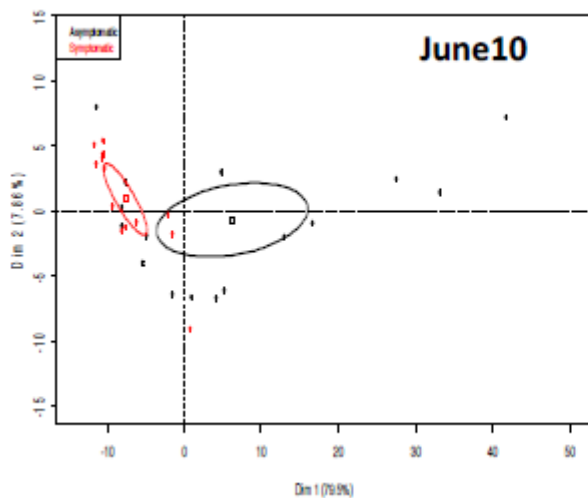
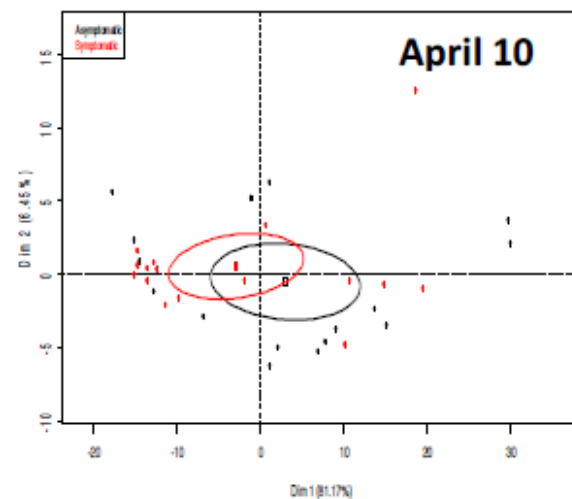
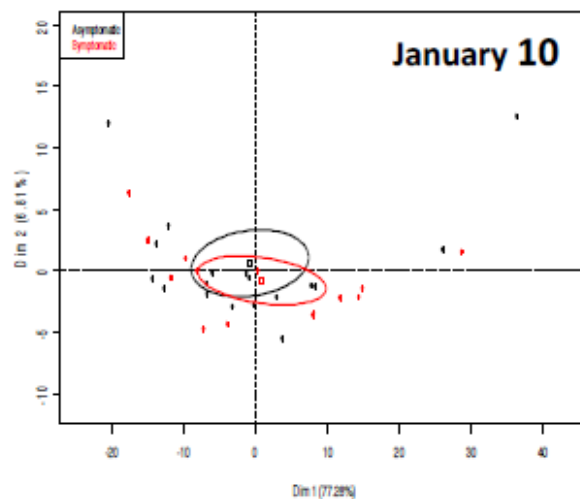
Analyse en composante principale de la distribution des populations fongiques dans des ceps **symptomatiques** et asymptotiques



Asymptomatique Symptomatique



Analyse en composante principale de la distribution des populations bactériennes dans des ceps **symptomatiques** et asymptomatiques



Comment identifie-t-on les micro-organismes dont on a obtenu la séquence 16S
Avantage de connaitre la séquence du génome de la vigne?

Conclusions

Outils très puissants mais qui déplacent la dépendance vers les outils informatiques, mathématiques et statistiques

Attention à la question biologique et au matériel expérimental

Phénotypage à haut débit? Reste limitant