Analyse et exploitation de séquences Sanger

Aurélie Canaguier, Aurélie Bérard, Isabelle Le Clainche, Aurelie Chauveau, Corinne Cruaud, Patricia Faivre Rampant

Analyse et exploitation de séquences Sanger

Formation des 11 et 12 mai 2017 – Evry

Aurélie Canaguier, Aurélie Bérard, Aurélie Chauveau, Isabelle Le Clainche
Corinne Cruaud, Patricia Faivre-Rampant

But :

- **Analyse des séquences Sanger produites par le Genoscope à partir de vos produits de PCR :**
  - Utilisation de Genalys pour voir, nettoyer et comparer les séquences
  - Détection de ***S**ingle **N**ucleotid **P**olymorphism* et/ou Insertion/Délétion
  - Production d'un fichier de sortie exploitable par d'autres outils

- **Exploitation de ces analyses :**
  - Interprétation des résultats obtenus en sortie de Genalys

Plan :

1) Pourquoi étudier le polymorphisme des génomes, description de polymorphismes des génomes et d'outils d'étude.

2) Description de polymorphismes nucléotidiques de type SNP et Insertion/Délétion et d'un outil d'étude.

3) Pas à pas avec Genalys.

4) Quelques pistes d'exploitation des SNP.

1.1) Pourquoi s'intéresse-t-on au polymorphisme des génomes ?

- **Améliorer la connaissance sur :**

  - La diversité génétique
  - La classification des espèces
  - L'évolution des espèces et leur capacité d'adaptation
  - Le fonctionnement du génome (notion d'allèles et d'expression différentielle)

- **Exploiter ces connaissances :**

  - Conservation éclairée des espèces
  - Caractérisation de fonctions physiologiques et gènes associés.
  - Sélection adaptée aux besoins de l'homme et de son environnement et création variétale

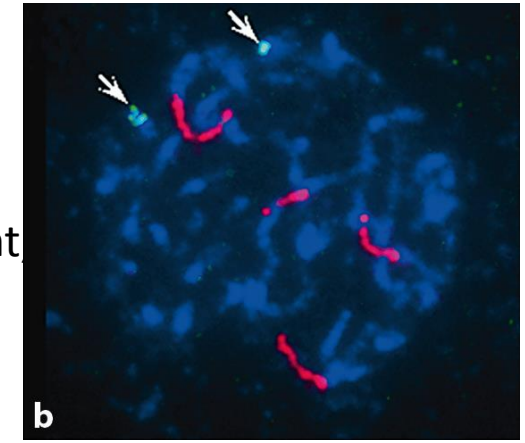
1.2) Quelques polymorphismes des génomes

- **Répétitions (duplication, triplication… n-plication)**
  - Grandes : régions chromosomiques entières
  - Petites : n répétitions de 2 à 6 bases (microsatellite), doublement d'un gène

- **Réarrangements** (inversions, Insertion / délétion ou « indel »)
  - Grandes : translocations chromosomiques, insertion ou délétion
  - Petites : n base(s) en plus ou en moins

- **Polymorphisme de structure au niveau d'un seul nucléotide** (SNP, *single nucleotide polymorphism*)
  - Résultat d'une substitution, erreurs lors de la réplication (Transitions ou transversions)
  - Résultat d'une mutation

1.3) Quelques outils d'étude des polymorphismes des génomes



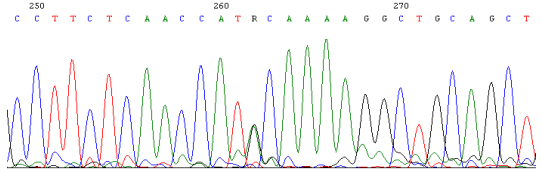
Détection de l'ADNr 18S-28S (rouge) et 5S (vert) par FISH sur des chromosomes de peuplier

• **Grandes répétitions et grands réarrangements chromosomique > La <u>cytogénétique</u> :**

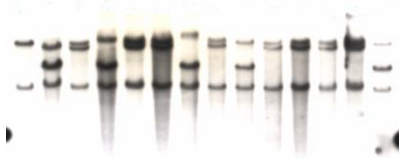
La coloration au Giemsa des chromosomes et l'étude des bandes claires et sombres qui apparaissent,
L'hybridation *in situ* de sondes marquées (fluorescence) sur les chromosomes (**FISH**),
sont des techniques très puissantes pour mettre en évidence ces évènements.

• **Petites répétitions, petits réarrangements, polymorphisme SNP**

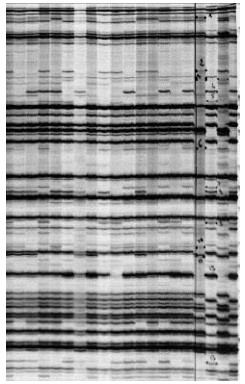
- Lyse enzymatique et dépôt sur gel (recherche de la présence/absence de sites enzymatiques) [RFLP]
- PCR aléatoire ou ciblée et dépôt sur gel (identification de polymorphisme de longueur, de la présence/absence de séquences)
- Séquençage et alignement sur une référence pour comparaison



SNP



RFLP
Restriction Fragment Length polymorphism



AFLP
Amplified Fragment Length Polymophism

2.1) Description des SNPs et Insertion/Délétion

Pour détecter des SNP, InDel, il faut comparer au moins 2 séquences correspondant à une même région d'un génome. L'une des séquences sera prise comme référence.

SNP et InDel ne présentent généralement que deux allèles et sont dits peu informatifs par rapport à d'autres marqueurs comme les microsatellites.

Mais ils sont très abondants, robustes et leur détection et exploitation sont faciles à automatiser.

2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAAGCCTGA(

Séquence ATGGTAAACCTGA(

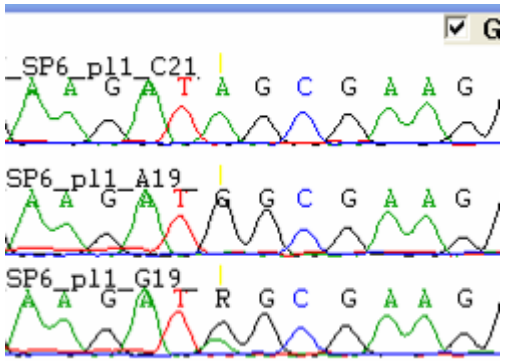
SNP

Sur un chromatogramme :

Homozygote 1

Homozygote 2

Hétérozygote



2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAAGCCTGA(

Séquence ATGGTAAACCTGA(

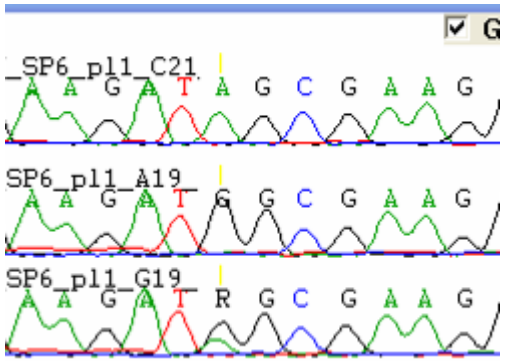
SNP

Sur un chromatogramme :

Homozygote 1

Homozygote 2

Hétérozygote

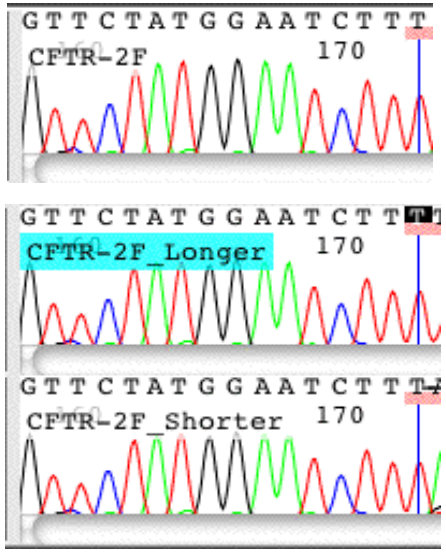


Référence TAGCGT-AT

Séquence TAGCGTCAT

indel

(Insertion/délétion)



2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAAGCCTGA(

Séquence ATGGTAAACCTGA(

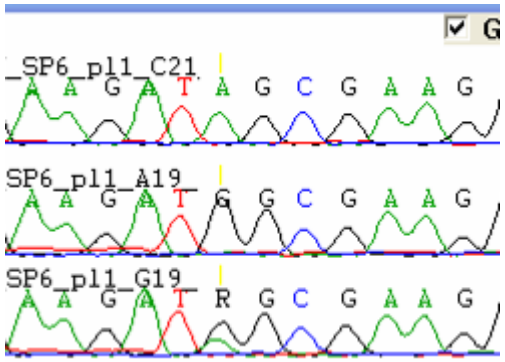
SNP

Sur un chromatogramme :

Homozygote 1

Homozygote 2

Hétérozygote



Référence TAGCGT-AT

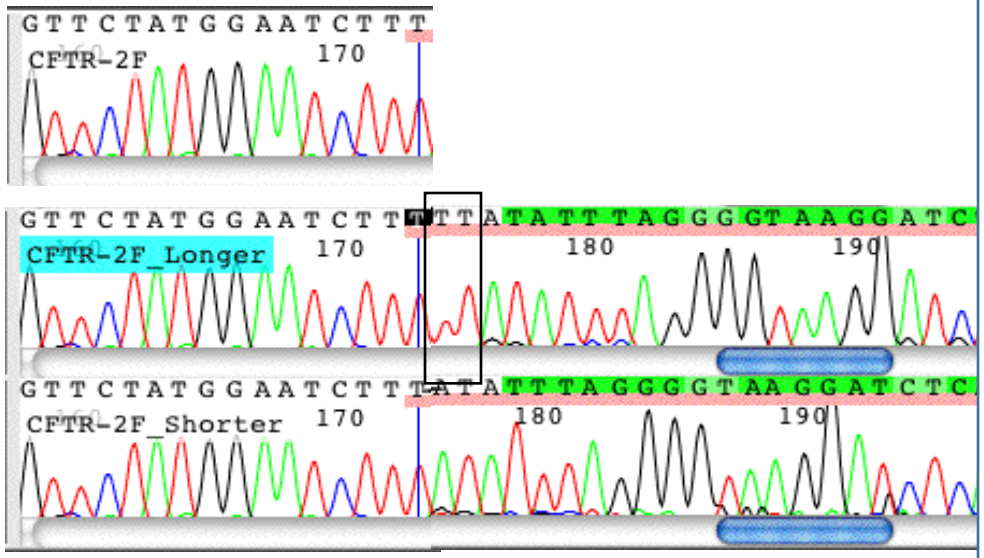
Séquence TAGCGTCAT

indel

(Insertion/délétion)

Homozygote 1

Homozygote 2



2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAAGCCTGA(

Séquence ATGGTAAACCTGA(

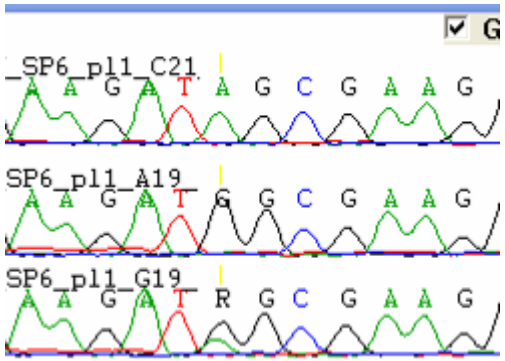
SNP

Sur un chromatogramme :

Homozygote 1

Homozygote 2

Hétérozygote



Référence 'TAGCGT-AT

Séquence 'TAGCGTCAT

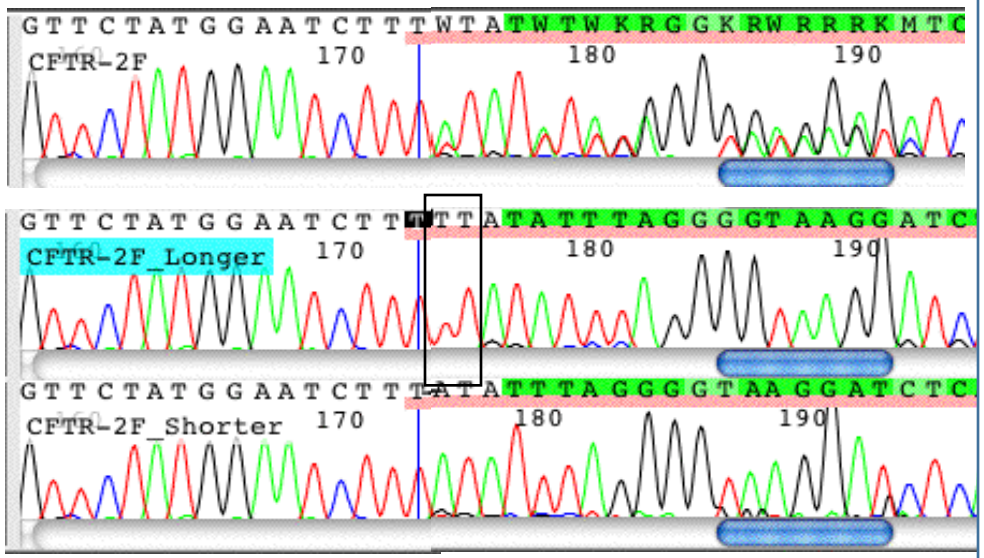
indel

(Insertion/délétion)

Hétérozygote

Homozygote 1

Homozygote 2



2.1) Description des SNPs et Insertion/Délétion

- **Pour étudier des SNP et InDel, il faut donc :**

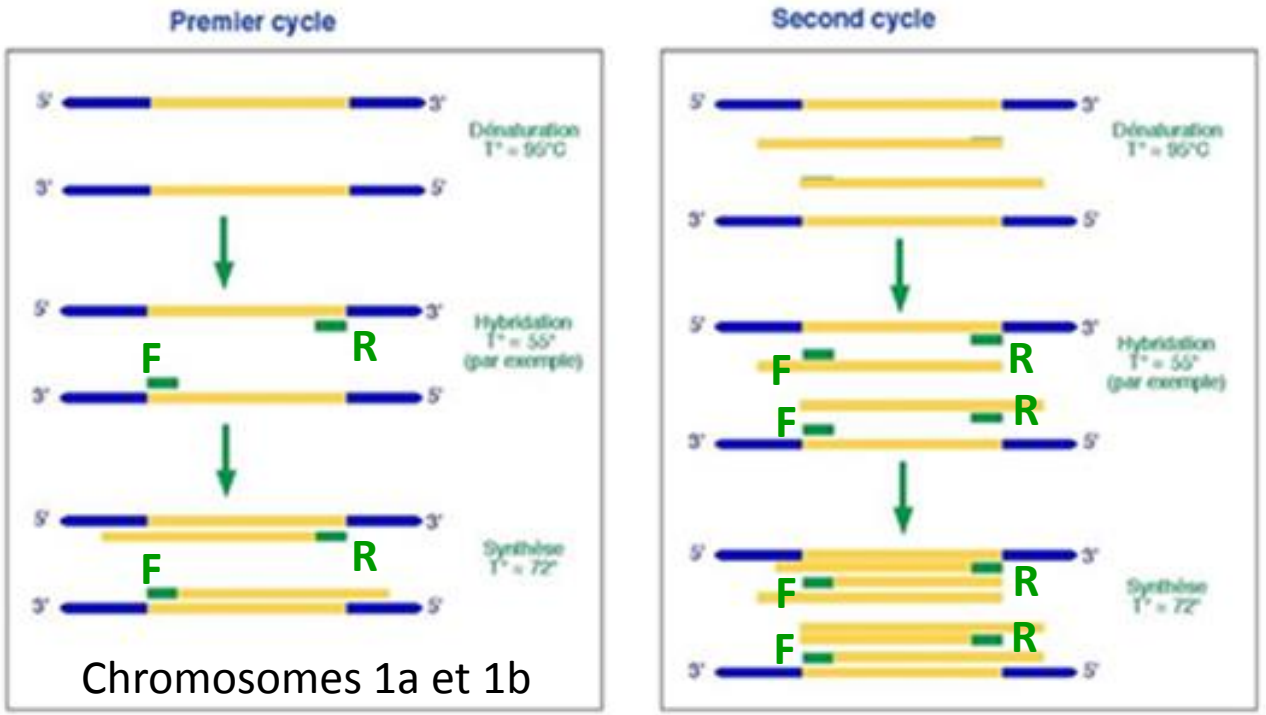
  - un outil pour produire les séquences que l'on veut comparer
  - Un outil pour comparer ces séquences et détecter les SNP, InDel
  - Un outil pour exploiter les SNP, InDel

2.2) Description d'un outil d'étude des SNPs et InDel : Le séquençage de fragments PCR

2.2) Description d'un outil d'étude des SNPs et InDel : Le séquençage de fragments PCR

- **La PCR sur génome diploïde homozygote**

L'ADN génomique est représenté en **bleu**.
Les amorces de PCR sont représentées en **vert (=primers ou oligonucléotides)**.
La région d'ADN génomique que l'on souhaite amplifier est représentée **en jaune**.



Chromosomes 1a et 1b

F = amorce Forward

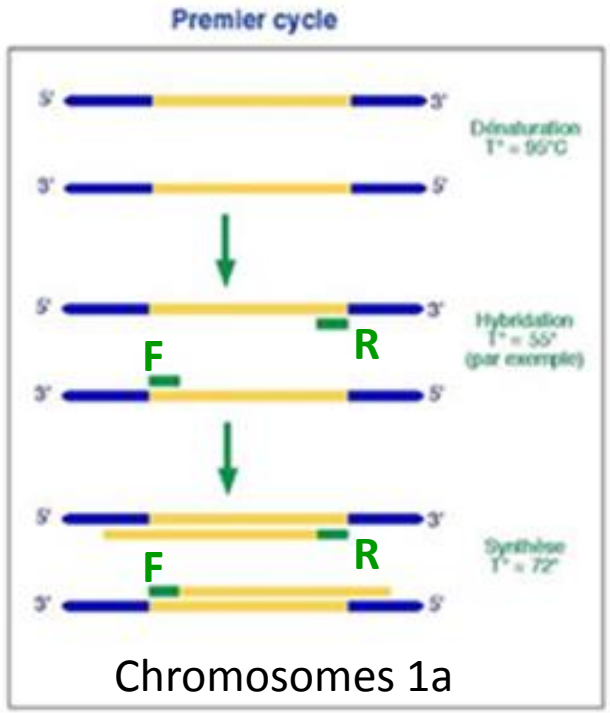
R = amorce Reverse

2.2) Description d'un outil d'étude des SNPs et InDel : Le séquençage de fragments PCR

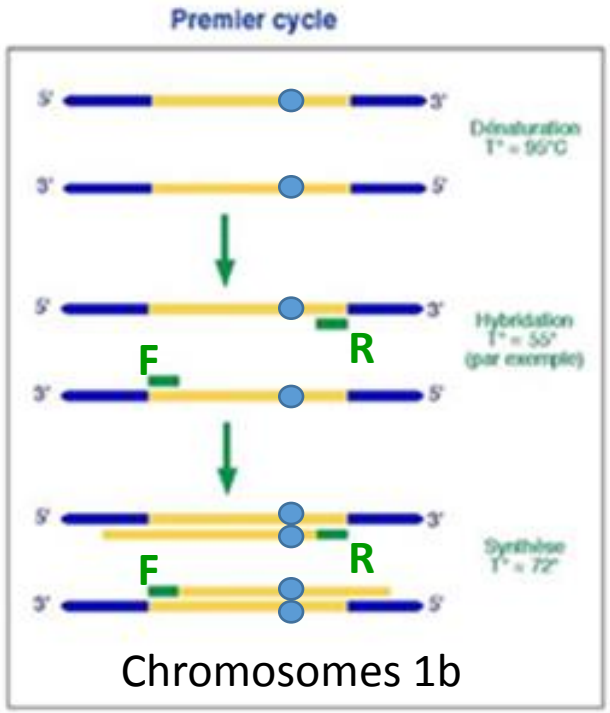
- **La PCR sur génome diploïde hétérozygote**

L'ADN génomique est représenté en **bleu**.
Les amorces de PCR sont représentées en **vert (=primers ou oligonucléotides)**.
La région d'ADN génomique que l'on souhaite amplifier est représentée **en jaune**.

- SNP ou In/Del



Chromosomes 1a



Chromosomes 1b

F = amorce Forward

R = amorce Reverse

2.2) Description d'un outil d'étude des SNPs et InDel : Le séquençage de fragments PCR

- **Le séquençage Sanger**

La lecture ordonnée de chacun des desoxyribonucléotides (dNTP) constituant le fragment de PCR.

Les 3 étapes principales sont :
1) Synthèse de brins de tailles aléatoires et différentes, terminées par un ddNTP fluorescent (un fluorochrome par type de base), à partir du fragment PCR.
2) Migration sur gel de ce mélange de brins et donc séparation de tous les brins (les petits brins migrent plus vites que les grands brins).
3) Lecture et enregistrement de la nature et de l'intensité de la fluorescence en fin de migration dans un fichier.ab1.

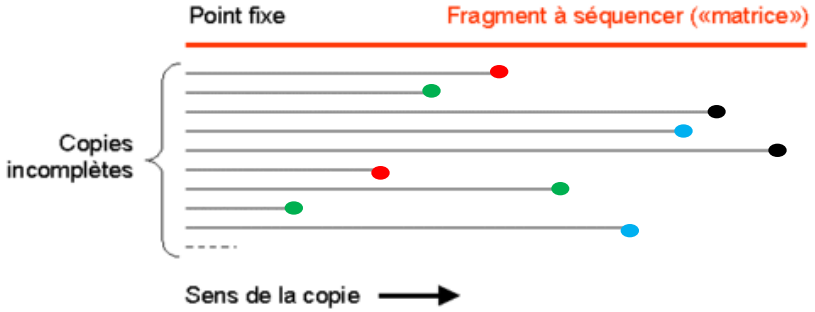
- **Le séquenceur**

C'est un **AB1** 3730 (Applied Biosystems ; L. Hood/M. Hunkapiller) équipé d'une nappe de 48 capillaires. Les produits de séquençage sont automatiquement pipetés à partir de plaques 96 puits.

2.2) Description d'un outil d'étude des SNPs et InDel : Le séquençage de fragments PCR

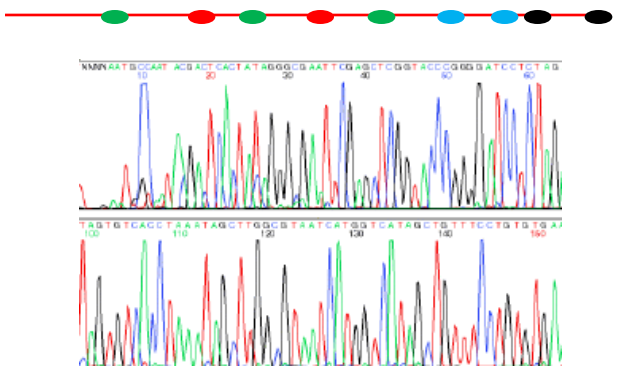
- **Le séquençage Sanger**

1) Synthèse



2) Migration dans un tube de verre très fin rempli de polymère (séquenceur capillaire)
=> séparation des brins.

3) Production d'**électrophorégrammes** enregistrés au format *.ab1.
Ce sont les données brutes.



3) Tutoriel pour l'alignement et la comparaison de séquences avec le logiciel Genalys

- **Genalys permet notamment :**
  - d'aligner des séquences,
  - de corriger les séquences,
  - caractériser les SNP et InDel,
  - Obtenir des fichiers de sortie exploitable de SNP et InDel.
- **Autres avantages :**
  - visualisation simultanée et interactive des séquences et chromatogrammes,
  - Intuitif et gratuit.
- **Inconvénient :**
  - Instable sous Windows 7 => il faut sauvegarder TRES régulièrement !!

- **Auteurs du logiciel** : Masazumi Takahashi & Fumihiko Matsuda
- Copyright © 1999-2001 Masazumi Takahashi
- Version Windows 2.8.3b
- Testé sur Windows XP (des problèmes sur Windows 7)

3.1) Installer Genalys

- **Copier l'archive 2.8.3b_GENALYS.zip**

- **Extraire tout le contenu de l'archive 2.8.3b_GENALYS.zip**

- **Créer un raccourci du fichier exécutable sur le bureau**

- **Auteurs du logiciel** : Masazumi Takahashi & Fumihiko Matsuda
- Copyright © 1999-2001 Masazumi Takahashi
- Version Windows 2.8.3b
- Testé sur Windows XP (des problèmes sur Windows 7)

3.2) Les fichiers de séquences

• **Les fichiers traces** sont visualisés de 2 manières :

- les **électrophorégrammes** (fichiers.ab1)
- la séquence (suite de nucléotides) obtenue avec un **logiciel de *base calling*** (« appel de base ») qui interprète chaque pic (nature et intensité) du fichier.ab1.

REMARQUE : **Genalys intègre cette fonction** de *base calling*. Il interprètera donc le fichier trace dès l'ouverture ! Le Genoscope utilise le logiciel Phred.

• **Les fichiers FASTA ou MULTIFASTA**

- Ce sont de simples fichiers textes, qui contiennent une ou plusieurs séquence(s).
- L'extension peut être « .fasta » ou « .fa» ou « .txt » ou « .seq ». Genalys lit très bien tous les fichiers avec ces extensions.
- Format caractéristique, pour chaque séquence :

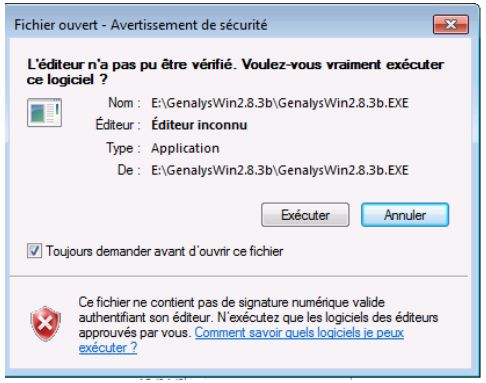
1ère ligne : > Nom_seq Identifiant ; Organisme….

n lignes de séquence (80 bases par ligne) utilisant les 4 lettres A, T, G, C ou les lettres du code IUPAC et « - » pour représenter une délétion.

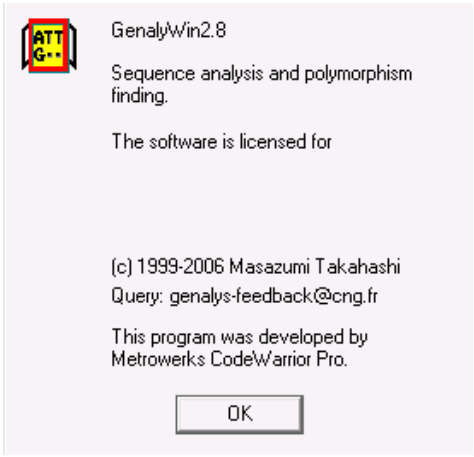
REMARQUE : En suivant ces règles, vous pouvez créer votre propre fichier de séquence.txt, Genalys pourra l'ouvrir !

3.3) Ouvrir des fichiers de séquences (1/6)

- **Ouvrir Genalys :** Double clic sur le raccourci du fichier exécutable

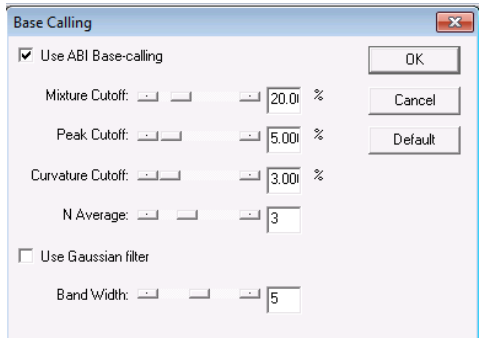


Exécuter



OK

- **Choisir le Base Calling :** Settings/Base Calling



OK

- **Choisir des séquences :** File/Enter Sequences…./

3.3) Ouvrir des fichiers de séquences (2/6)

Si le logiciel ne veut plus insérer de nouvelles séquences :
- Aller dans le répertoire où sont stockées les séquences,
- Faire glisser celle de son choix dans la fenêtre Genalys.

- **Choisir des séquences :** File/Enter Sequences…./

  Un échantillon a 2 séquences (Forward et Reverse)
  Fermer la fenêtre Project Info
- **Choisir la séquence de référence :** File/Enter Sequences…./

  La faire glisser tout en haut avec la souris.

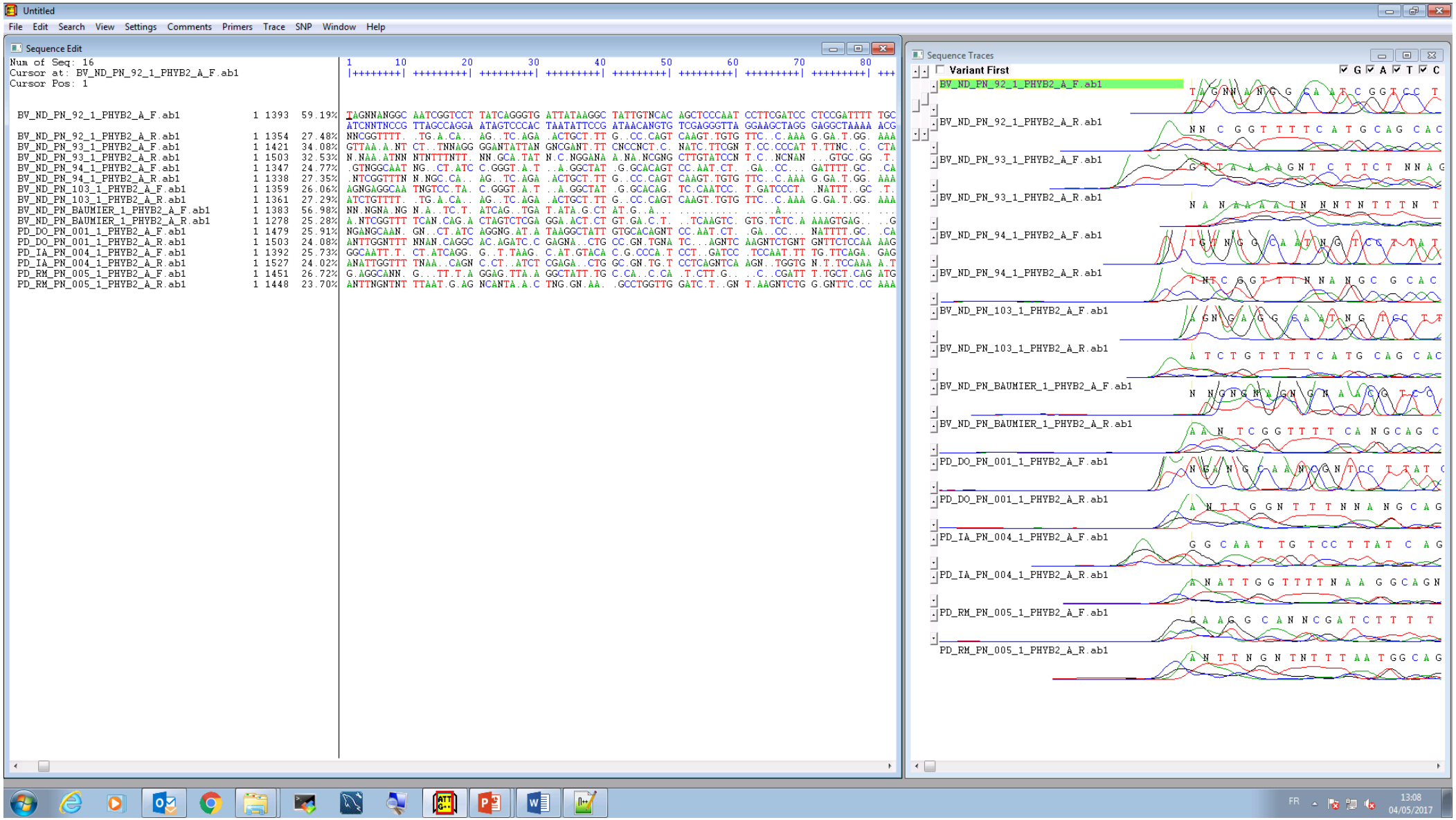
- **Agencer les fenêtres Genalys** pour voir un maximum de séquences

- **TEST**

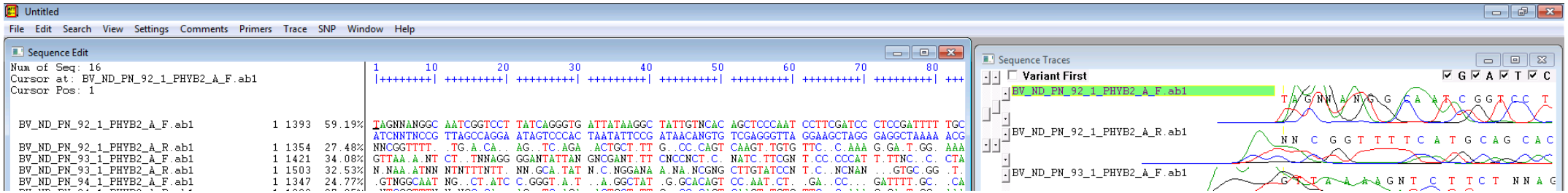
  Charger les séquences du gène PHYB2 du répertoire SEQ_TEST/

  Charger la séquence de référence du gène PHYB2 Seq_de_ref_Genome_a_Lecole/Seq_ref_PHYB2_P_nigra_020415.txt

3.3) Ouvrir des fichiers de séquences (3/6)



3.3) Ouvrir des fichiers de séquences (4/6)



- **Deux fenêtres se sont ouvertes :**
  - Sequence Edit : séquence(s) lue(s) à partir des fichier.ab1
  - Sequence Traces : chromatogramme(s) correspondant

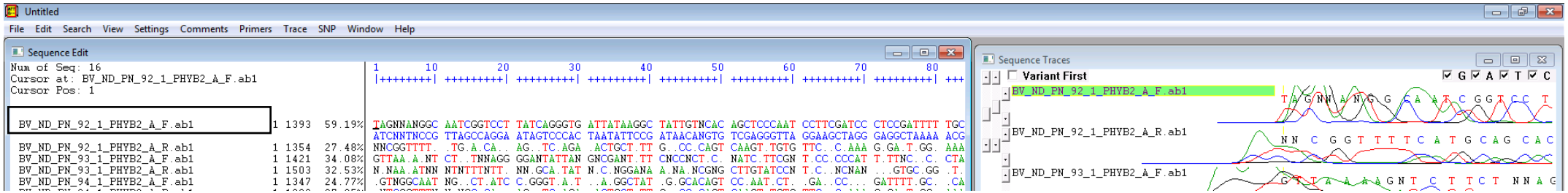
En ouvrant un fichier.txt, seule la fenêtre « Sequence Edit » sera informée

- **Sequence Edit** , elle est composée de 2 parties :

A gauche

- Le nombre de séquences.
- La position du curseur.
- Le nom des séquences, le numéro de la 1ère base, le numéro de la dernière base, le pourcentage d'identité entre la séquence et la séquence de référence.
- Les séquences sont classées par ordre alphabétique les unes en dessous des autres.

3.3) Ouvrir des fichiers de séquences (5/6)



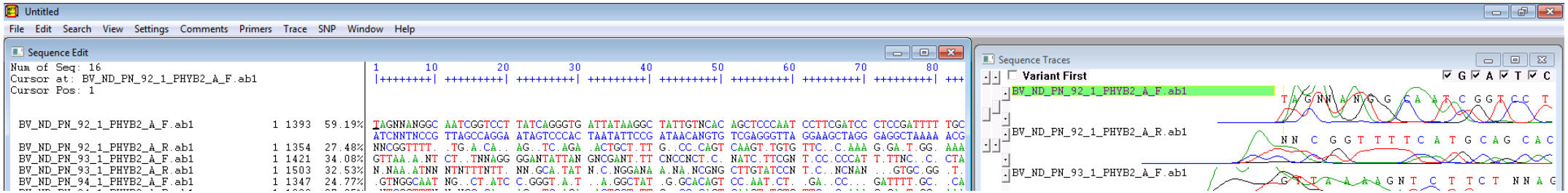
En ouvrant un fichier.txt, seule la fenêtre « Sequence Edit » sera informée

- **Sequence Edit** , elle est composée de 2 parties :

A droite

- La séquence où chaque base est associée à une couleur.
- Une échelle en nombre de base est indiquée en haut (cliquer sur l'échelle pour repositionner l'origine).
- La première séquence de la liste est la **séquence de référence.** La séquence complémentaire est affichée en bleu juste en dessous. En remontant une séquence tout en haut, vous en faites la nouvelle séquence de référence.
- Vous pouvez sélectionner une base dans une séquence, qui est alors surlignée en vert et la position du curseur est indiquée à gauche.
- Avec un clic droit sur la base pré-sélectionnée, la fenêtre « Sequence Trace » se synchronise en centrant la base.

3.3) Ouvrir des fichiers de séquences (6/6)



- **Sequence Trace** :

En ouvrant un fichier.txt, seule la fenêtre « Sequence Edit » sera informée

- Vous pouvez faire varier la largeur des pics de l'ensemble des fichiers traces à l'aide du premier « ascenseur » à gauche.
- Vous pouvez faire varier la hauteur des pics de l'ensemble des fichiers traces, ou de chaque fichier, avec les ascenseurs suivants.
- Vous pouvez faire disparaitre les pics correspondant à l'une des quatre bases A, T, G, C en décochant les boites en haut à droite (Analyse de régions confuses).

- **Ajout / Suppression de séquence(s)** :

(+) Faire glisser les séquences désirées dans l'une ou l'autre des deux fenêtres de Genalys, depuis votre répertoire.

(-) Sélectionner les séquences à supprimer au niveau de leurs noms (Touche Ctrl pour une sélection multiple)

(-) Faire glisser la sélection à l'extérieur de la fenêtre « Sequence Edit »

3.4) Aligner des séquences et commencer l'analyse

- **Alignement** :
  - Placer le curseur sur la séquence de référence pour que Genalys aligne toutes les séquences.
  - Search/Locate By Similarity ou Search/Match to Display (alignement des sequences dans la fenêtre)
  - Le nom des séquences Forward et non alignées : en noir
  - Le nom des séquence Reverse : en vert (elles sont réverse-complémentées).
  - Lorsque la séquence est identique à la référence il y a un « . » à la place du A,T,G,C ou N

=> En se baladant le long de la séquence avec l'ascenseur de la fenêtre « Sequence Edit », on voit les incohérences entre nos séquences et la séquence de référence.

3.4) Aligner des séquences et commencer l'analyse

- **Alignement** :
  - Placer le curseur sur la séquence de référence.
  - Search/Locate By Similarity
  - Le nom des séquences Forward et non alignées : en noir
  - Le nom des séquence Reverse : en vert (elles sont réverse-complémentées.
  - Lorsque la séquence est identique à la référence il y a un « . » à la place du A,T,G,C ou N

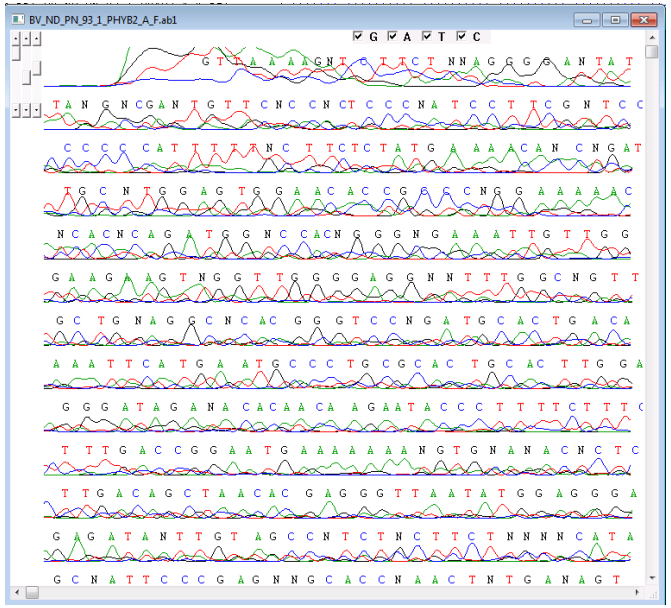
=> En se baladant le long de la séquence avec l'ascenseur de la fenêtre « Sequence Edit », on voit les incohérence entre nos séquences et la séquence de référence.

- **Analyse – Suppression des séquences de mauvaise qualité** :
  - Dans la fenêtre « Sequence Edit », double-clic sur une séquence mal alignée pour voir le chromatogramme associé

BV_ND_PN_92_1_PHYB2_A_F.ab1
BV_ND_PN_92_1_PHYB2_A_R.ab1

  - Pourcentage d'identité < 50%

⇒ Suppression de la séquence (cf diapo 21)
⇒ Mettre la séquence à 0 :
Sélection multiple
Edit/Zero Sequence
Déplacer ces séquences en bas



3.4) Aligner des séquences et commencer l'analyse

- **Analyse – Localisation des amorces**
  - Ouvrir le fichier amorces_PHYB2.txt.
  - Copier la sequence PHYB2_F_test
  - Dans la fenêtre "Sequence Edit", mettre le curseur sur le nom de la reference
  - Search/Sequence Search…
  - Coller la sequence PHYB2_F
  - Clic droit au-dessus de la séquence surlignée et ajouter un commentaire.

```
>PHYB2_F_test
TGTTGGTCAGGATGTTACAGGTC
>PHYB2_R_test
CAAGATCAATGTCTCGTATG
```

=> La base 1 de la séquence de référence est la base 1 de l'amorce Forward

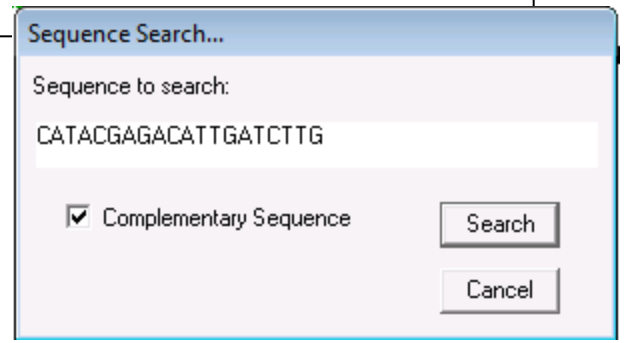
3.4) Aligner des séquences et commencer l'analyse

- **Analyse – Localisation des amorces**
  - Ouvrir le fichier amorces_PHYB2.txt.
  - Copier la sequence PHYB2_F_test
  - Dans la fenêtre "Sequence Edit", mettre le curseur sur le nom de la reference
  - Search/Sequence Search…
  - Coller la sequence PHYB2_F et « Search »
  - Clic droit au-dessus de la séquence surlignée et ajouter un commentaire.

```
>PHYB2_F_test
TGTTGGTCAGGATGTTACAGGTC
>PHYB2_R_test
CAAGATCAATGTCTCGTATG
```

=> La base 1 de la séquence de référence est la base 1 de l'amorce Forward

- Copier la sequence PHYB2_R_test
- Dans la fenêtre "Sequence Edit", mettre le curseur sur le nom de la reference
- Search/Sequence Search
- Coller la sequence PHYB2_R_test, sélectioner **Complementary Sequence** et "Search"
- Clic droit au-dessus de la séquence surlignée et ajouter un commentaire.



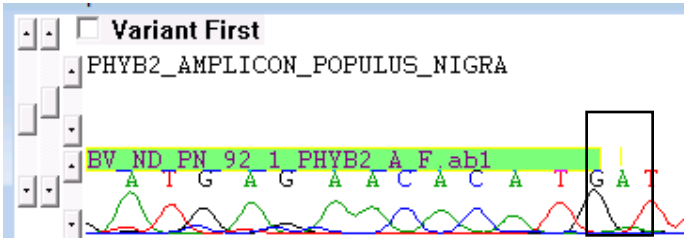
3.4) Aligner des séquences et commencer l'analyse

- **Analyse – Suppression de séquences inutiles**
  - Sélectionner la séquence en amont de la position 1 et supprimer (backsupp)
  - Sélectionner la séquence en aval de la position 663 et supprimer (backsupp)

- **Analyse – Nettoyage et amélioration de l'alignement**

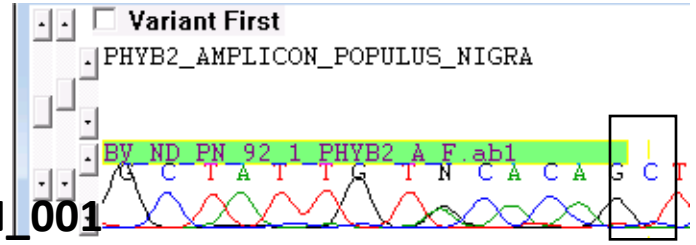
1

- **Pos_132** : clic droit sur la séquence de **BV_ND_PN_92_1**
- Supprimer la base en trop : double clic sur cette base dans la fenêtre « Sequence Trace »
- Faire la même chose pour **PD_DO_PN_001** et **BV_ND_PN_BAUMIER**



2

- **Pos_84** : clic droit sur la séquence de **BV_ND_PN_92_1**
- Supprimer la base en trop.
- Faire la même chose pour **BV_ND_PN_103** et **BV_ND_PN_BAUMIER** et **PD_DO_PN_001**



3

- **Pos_76** : clic droit sur la séquence de **BV_ND_PN_BAUMIER**
- Supprimer la base en trop.

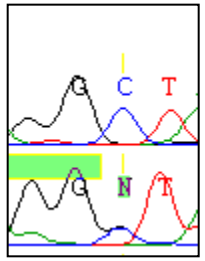
4

- **Regarder** les chromatogrammes en amont de la position 70
- **Sélectionner** (en se mettant à l'extrémité gauche de la fenêtre **et supprimer** les séquences en amont de la position 70 (backsupp)

3.4) Aligner des séquences et commencer l'analyse

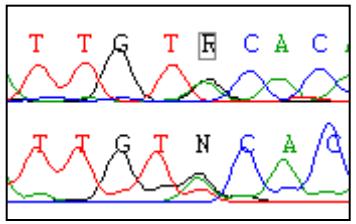
- **Analyse – Nettoyage et amélioration de l'alignement**

5 - **Pos_72** : C **BV_ND_PN_92_1_F** et S **BV_ND_PN_92_1_R** => incohérence => correction :
Dans la fenêtre « Sequence Trace » sélectionner le N et le remplacer par un C
Faire la même chose pour **BV_ND_PN_94_1_R** et **BV_ND_PN_103_1_R**



6 - **Pos_74** : N **PD_IA_PN_004_1_R** => correction : A

7 - **Pos_79** : clic droit sur la séquence de **BV_ND_PN_92_1_F** et **_R** : R
Faire la même chose pour **BV_ND_PN_94_1_F** et **_R**
Faire la même chose pour **BV_ND_PN_103_F** et **_R**



8 - **Pos_83** : C **BV_ND_PN_92_1_F** et S **BV_ND_PN_92_1_R** => incohérence => correction :
Dans la fenêtre « Sequence Trace » sélectionner le N et le remplacer par un A
Faire la même chose pour **BV_ND_PN_94_1_R** et **BV_ND_PN_103_1_R**

9 - **Pos_97** : **BV_ND_PN_BAUMIER** a un beau SNP

3.4) Aligner des séquences et commencer l'analyse

- **Analyse – Nettoyage et amélioration de l'alignement**

10 - **Pos_106** : clic droit sur la séquence de **BV_ND_PN_92_1_F** et **_R** : R
Faire la même chose pour **BV_ND_PN_94_1_F** et **_R**
Faire la même chose pour **BV_ND_PN_103_F** et **_R**
Faire la même chose pour **PD_DO_PN_001_1_F** et **_R**

11 - **Pos_128, Pos_130, Pos_149, Pos_155, Pos_167, Pos_169, Pos_179, Pos_224, Pos_247** et **Pos_266** : corriger les incohérences : N => A

12 - Jusqu'à la position **Pos_460 :** incohérences à corriger

13 - **Pos_462** : SNP **BV_ND_PN_BAUMIER**

14 - **Pos_481** et **Pos_497** : N à supprimer (double clic sur cette base dans la fenêtre « Sequence Trace ».

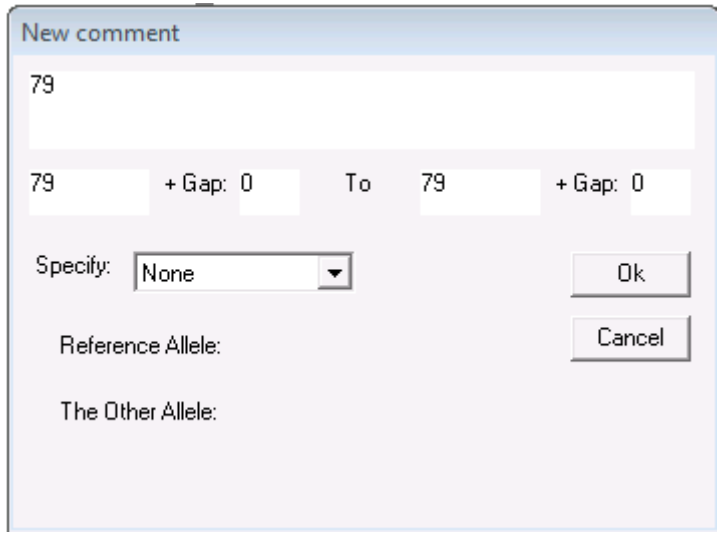
15 - **Pos_545, Pos_554** et **Pos_556** : C/A/N à supprimer

16 - **A partir de Pos_571** : Les chromatogrammes sont décalés=> sélection des séquences et suppression

3.4) Aligner des séquences et commencer l'analyse

- **Analyse – Edition des SNP – version 250pb**

17 - **Pos_79** : sélectionner la base de la référence où il y a le SNP
Clic droit au-dessus de cette base
Une fenêtre « New comment » s'ouvre.
Specify/SNP
The Other Allele: A
By Combination : R et cliquer sur « OK »



18 - **Pos_97 :** SNP G/A

19 - **Pos_106** : SNP G/A

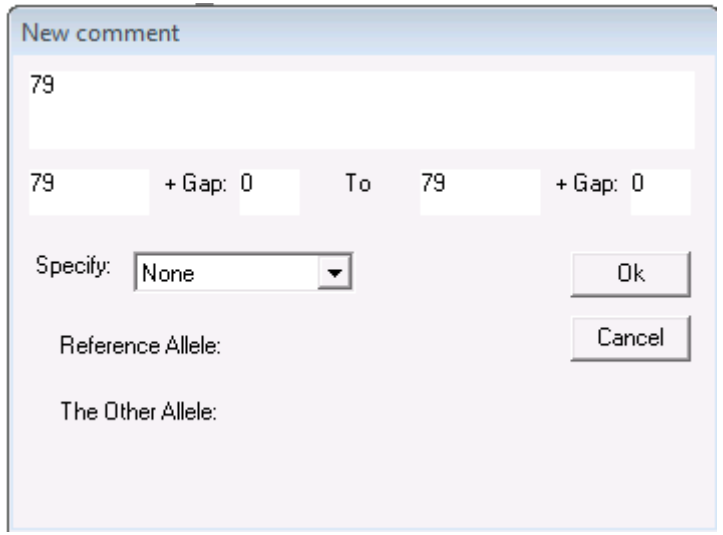
20 - Sélection des séquences de 251pb à la fin et suppression (backsupp)

21 - File/Save SNP => on obtient un fichier Excel avec les informations des 3 SNP pour les 7 individus (14 séquences)

3.4) Aligner des séquences et commencer l'analyse

- **Analyse – Edition des SNP – version 570pb**

17
- **Pos_79** : sélectionner la base de la référence où il y a le SNP
  Clic droit au-dessus de cette base
  Une fenêtre « New comment » s'ouvre.
  Specify/SNP
  The Other Allele: A
  By Combination : R et cliquer sur « OK »



18 - **Pos_97 :** SNP G/A

19 - **Pos_106** : SNP G/A

20 - **Pos_462** : SNP A/C

21 - Sélection des séquences de 571pb à la fin et suppression (backsupp)

22 - File/Save SNP => on obtient un fichier Excel avec les informations des 4 SNP pour les 7 individus (14 séquences)

4) Exploiter les SNP

- **Créer un multifasta des séquences**

  - Ouvrir le fichier PHYB2_250pb_SNP.xls avec Excel

Ligne 1 : nom du projet

Lignes 2 à 6 : description des commentaires (amorces et SNP)

PHYB2_250pb.gpr	Name_comment	START_comment	END_comment	Allele_Ref	Allele_Alternatif	Séquence
	PHYB2_F	1	23			
SN	79	79	79	G	A	TGTTGGTCA
SN	97	97	97	G	A	TGTTGGTCA
SN	106	106	106	G	A	TGTTGGTCA
	PHYB2_R	643	662			

Lignes 8 à 17 : code de chaque génotype

SNP	Code
X	0
Y	1
X/Y	2
XXX/Y	3
X/YYY	4
X/Y?	5
X/Y/Z	6
N	-1

X = Reference, Y = Sample

Lignes 19 à 40 : description de chaque séquence

4) Exploiter les SNP

- **Créer un multifasta des séquences**

  - Ouvrir une nouvelle feuille et la nommer « multifasta »
  - Sélectionner et copier les cellules B24 à K46
  - Coller ces cellules en A1 de la feuille multifasta
  - Insérer une ligne entre chaque séquence
  - Vider les colonnes B à I
  - Sélectionner les cellules de la colonne J, couper et la coller en B2
  - En C1 écrire la formule : =CONCATENER(">";A1)
  - En C2 écrire la formule : =B2
  - Sélectionner C1 et C2 et étirer ces 2 cellules simultanément.
  - Sélection de la colonne C et la copier en valeur dans la colonne D

4) Exploiter les SNP

- **Identification du genre de peuplier séquencé / Estimer une distance entre les séquences**

- Ajouter les séquences de références disponibles dans votre projet Genalys
- Identifier, commenter et exporter les SNP
- Créer un multifasta des séquences
- Coller les séquences en entrée du site :

Si vous n'avez qu'un ou peu ou plusieurs d'individu(s) séquencé(s)

http://www.phylogeny.fr/index.cgi

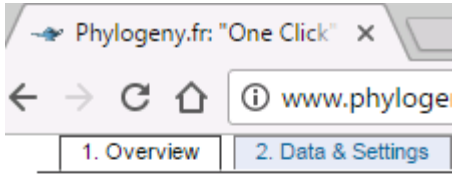
1.Dereeper A., Audic S., Claverie J.M., Blanc G. *BLAST-EXPLORER helps you building datasets for phylogenetic analysis.* BMC Evol Biol. 2010 Jan 12;10:8. (PubMed)
2.Dereeper A.*, Guignon V.*, Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. *Phylogeny.fr: robust phylogenetic analysis for the non-specialist.* Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19. (PubMed) *: joint first authors
3.Edgar RC. *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res. 2004, Mar 19;32(5):1792-7. (PubMed)
4.Castresana J. *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Mol Biol Evol. 2000, Apr;17(4):540-52. (PubMed)
5.Guindon S., Gascuel O. *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol. 2003, Oct;52(5):696-704. (PubMed)
6.Anisimova M., Gascuel O. *Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative.* Syst Biol. 2006, Aug;55(4):539-52. (PubMed)
7.Chevenet F., Brun C., Banuls AL., Jacq B., Chisten R. *TreeDyn: towards dynamic graphics and annotations for analyses of trees.* BMC Bioinformatics. 2006, Oct 10;7:439. (PubMed)

4) Exploiter les SNP

• **Identification du genre de peuplier séquencé / Estimer une distance entre les séquences**

http://www.phylogeny.fr/index.cgi

- Obtention d'un arbre phylogénétique
- Phylogény Analysis/ « One Click »
- Name of the analysis : PHYB2
- Coller la colonne D du fichier Excel (séquences au format multifasta)
- Mettre son adresse mail pour recevoir le lien des résultats de l'analyse



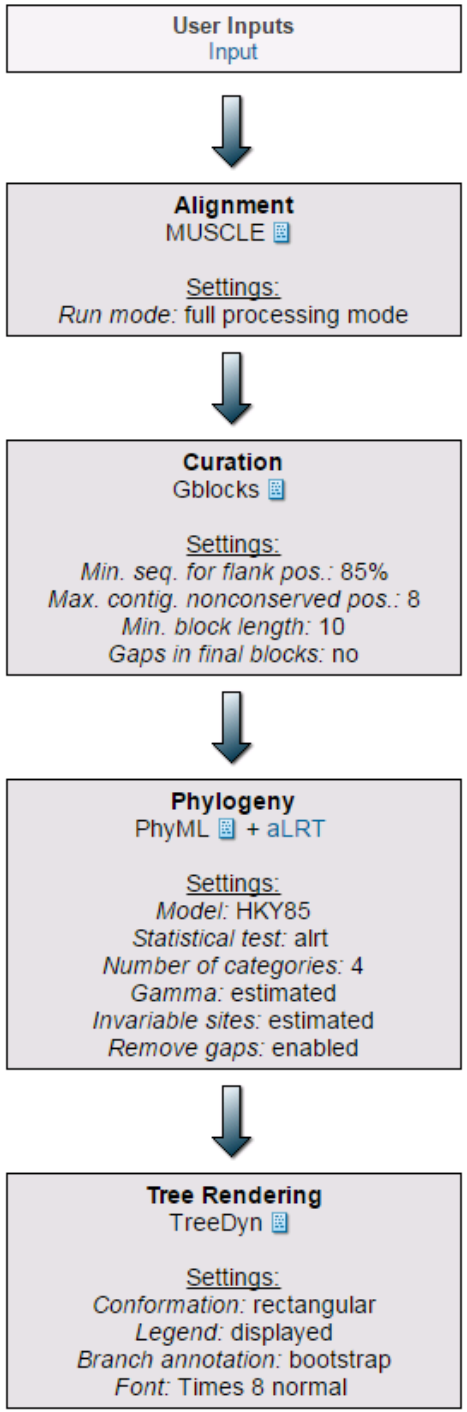
Create a new workflow

The analysis was performed on the Phylogeny.fr platform and comprised the following steps.

Process Info:
Number of taxa: 15
Average sequences length: 231
(between 200 and 663)
Assumed to be: DNA

Computation:

Overall time: 0 seconds



4) Exploiter les SNP

- **Identification du genre de peuplier séquencé / Estimer une distance entre les séquences**

http://www.phylogeny.fr/index.cgi

- Obtention d'un arbre phylogénétique

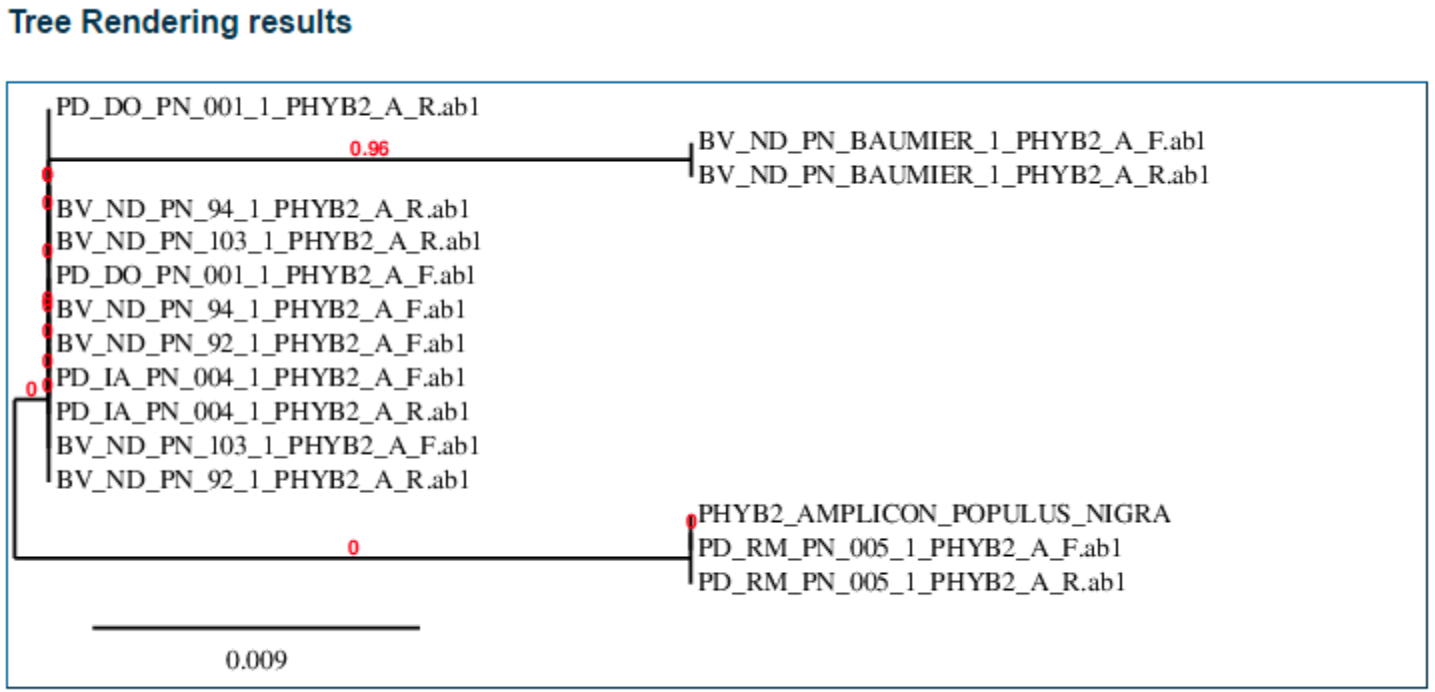


Figure 1: *Phylogenetic tree.*

4) Exploiter les SNP

- **Identification du genre de peuplier séquencé / Estimer une distance entre les séquences**

- Cliquer sur Reroot (outgroup) puis sur les séquences BV_ND_PN_BAUMIER

http://www.phylogeny.fr/index.cgi

Select an action and click leaf or internal branch:

Colorize ☑ leaf ☑ branch *choose a color* blue *and a legend label*

Reroot (**outgroup**)

Flip (flip an entire tree at a node)

Tree Rendering results

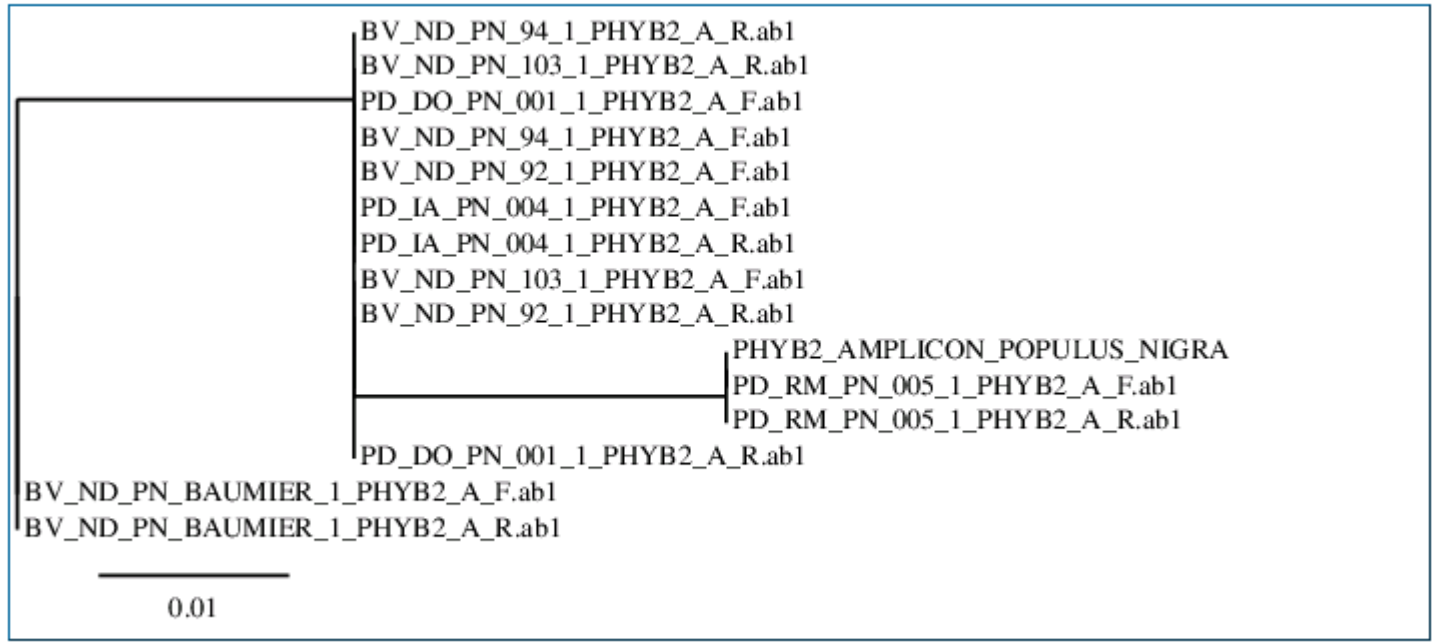


Figure 1: *Phylogenetic tree.*

4) Exploiter les SNP

- **Identification du genre de peuplier séquencé / Estimer une distance entre les séquences**

- Si vous avez un individu, vous verrez avec quelle référence il sera associé
- Si vous avez plusieurs individus, vous pourrez peut-être associer un groupe à une origine géographique ou un phénotype…

4) Exploiter les SNP

- **Identification du genre de peuplier séquencé / Estimer une distance entre les séquences**

- Si vous avez un individu, vous verrez avec quelle référence il sera associé
- Si vous avez plusieurs individus, vous pourrez peut-être associer un groupe à une origine géographique ou un phénotype…

- **Caractérisation du SNP : a-t-il un impact sur la protéine ?**

Populus trichocarpa v3.0

- Aller sur le site du JGI :
  https://phytozome.jgi.doe.gov/pz/portal.html# Tools/BLAST

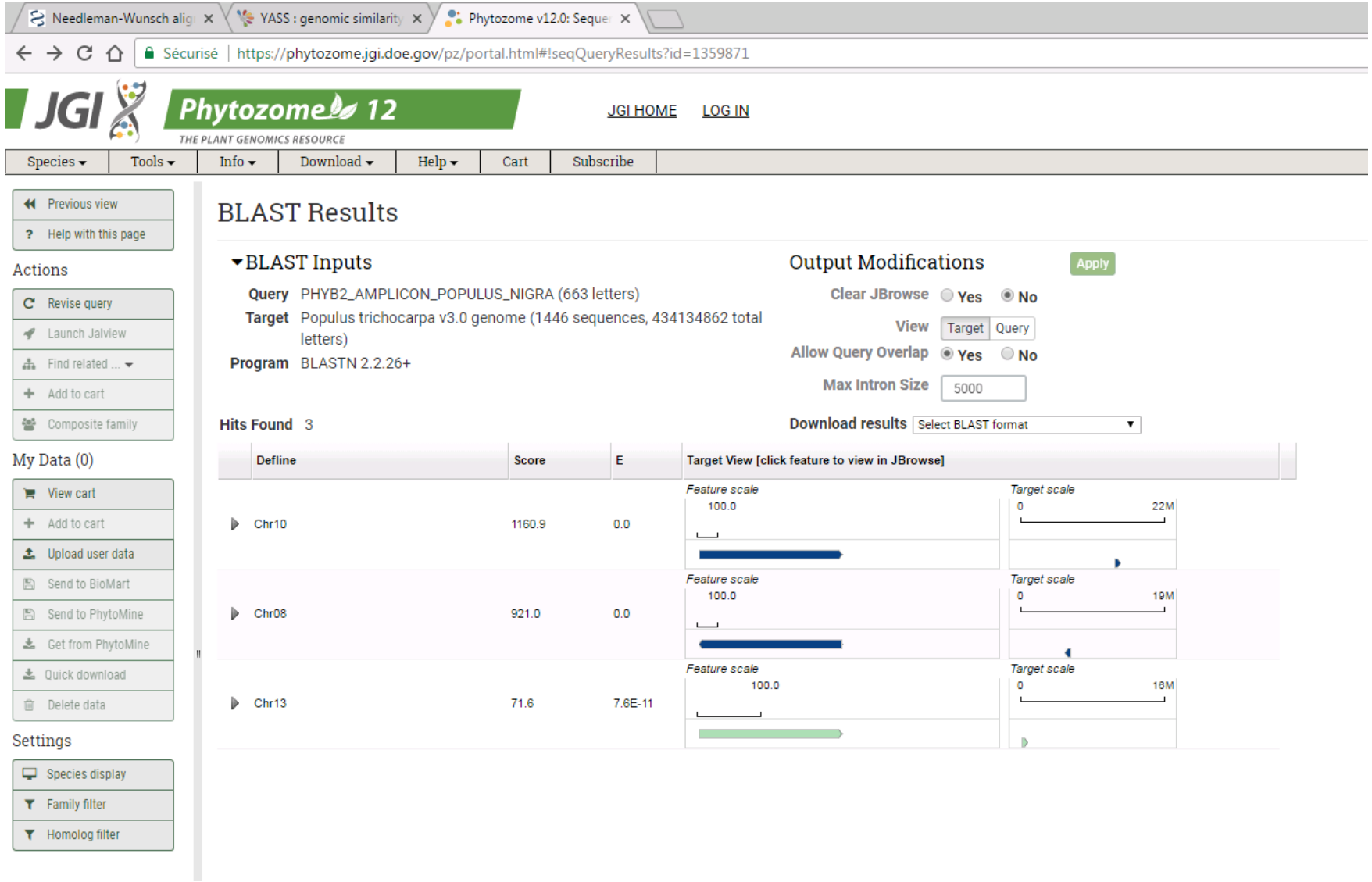
- 1.Select a Target / Populus trichocarpa v3.0

- 2.Build your query / Coller la séquence de la référence >PHYB2_AMPLICON_POPULUS_NIGRA et GO

4) Exploiter les SNP

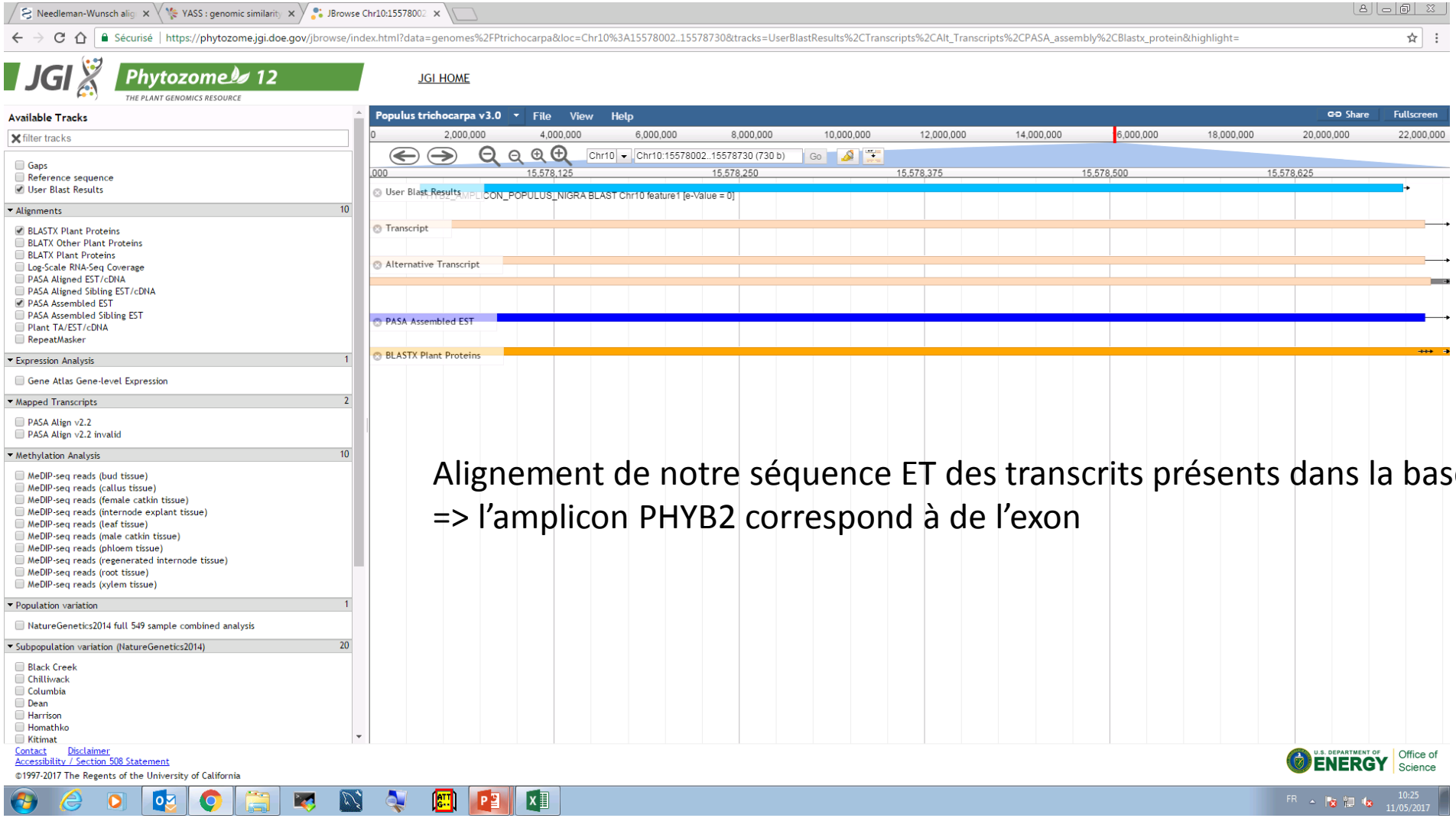
- **Caractérisation du SNP : a-t-il un impact sur la protéine ?**

- Cliquer sur le meilleur résultat



4) Exploiter les SNP

- **Caractérisation du SNP : a-t-il un impact sur la protéine ?**



Alignement de notre séquence ET des transcrits présents dans la base
=> l'amplicon PHYB2 correspond à de l'exon

4) Exploiter les SNP

- **Caractérisation du SNP : a-t-il un impact sur la protéine ?**

- Recherche du codon START (méthionine ATG) dans la séquence de référence PHYB2_AMPLICON_POPULUS_NIGRA

TGTTGGTCAGGATGTTACAGGTCAAAAAGTGGTAATGGACAAATACGTCCTTATACAAGGTGATTATAAGGCT
ATTGTGCACAGTCCCAATCCTTCGATCCCTCCGATTTTTGCTTCAGATGAGAACACATGTTGCTTGGAGTGGAA
CACTGCCATGGAAAAACTCACAGGATGGTCCAGGGGCGAAGTTGTTGGGAAGATGTTGGTTGGGGAGGTTTT
TGGCAGTTGCTGTAGGCTCAAGGGTCCAGATGCACTGACAAAATTCATGATTGCCCTGCACAATGCAATTGGA
GGGATAGATACAGACAAGTTACCCTTTTCATTCTTTGACCGGAATGAAAAAAATGTGCAAACTCTCTTGACAG
CTAACAAGAGGGTTAATATGGAGGGAGATATTATTGGAGCCTTCTGCTTCTTGCAGATTGCAATTCCTGAGTT
GCAGCAAACTTTGAAAGTTCAGAAACAGCAGGAAAAAAAAAAACTTTGCAAGGATGAAAGAGTTGGCTTACAT
TTGCCAGGAAATAAAAAATCCTTTGAGTGGTATACGCTTTACCAACTCACTTTTGGAGAACACAGACTTGACTG
AGGATCAACAGCAGTTTCTCGAGACTAGTGCTGCATGTGAAAAACAGATATTGAAGATCATACGAGACATTG
ATCTTGA¶

- Sous Genalys, afficher la traduction des codons en acides aminés : View / Show Translation
- Choisir le bon cadre de lecture (il ne produit pas de codon stop dans la séquence) : View / Frame / 1, 2 ou 3

=> Vous obtenez le START (1ère méthyonine)

- **Pos_79** : GTG = Val=Valine versus GTA = Valine
- **Pos_97** : TCG = Ser=Serine versus TCA = Serine
- **Pos_106** : CCG = Pro=Proline versus CCA = Proline

4) Exploiter les SNP

- **Caractérisation du SNP : a-t-il un impact sur la protéine ?**

- Le SNP détecté dans le fragment de 250pb de PHYB2 n'a aucun impact sur la protéine

=> on parle de polymorphisme neutre

- Il existe des mutations favorables, délétères et neutres

Merci et Bonnes Analyses !