



**HAL**  
open science

## Analyse et exploitation de séquences Sanger

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

► **To cite this version:**

Aurélié Canaguier, Aurélié Bérard, Isabelle Le Clainche, Aurelie Chauveau, Corinne Cruaud, et al..  
Analyse et exploitation de séquences Sanger. École thématique. France. 2017. hal-03507412

**HAL Id: hal-03507412**

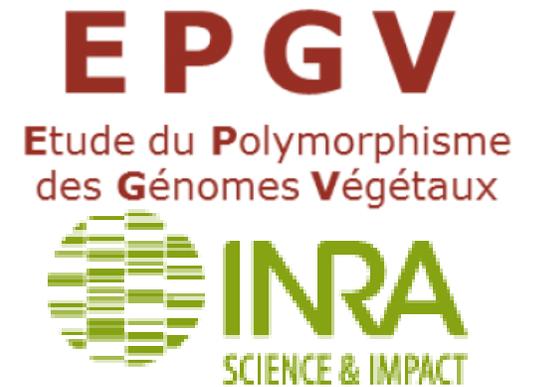
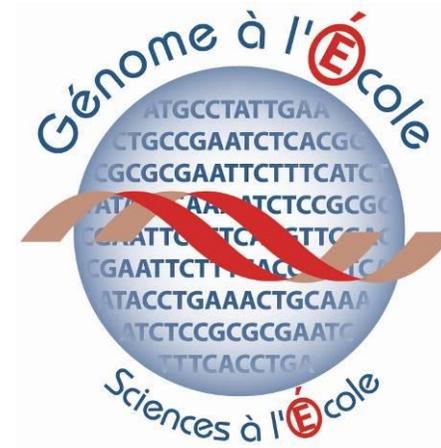
**<https://hal.inrae.fr/hal-03507412>**

Submitted on 3 Jan 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Analyse et exploitation de séquences Sanger



**Formation des 11 et 12 mai 2017 – Evry**

Aurélié Canaguier, Aurélié Bérard, Aurélié 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 *Single Nucleotid Polymorphism* 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

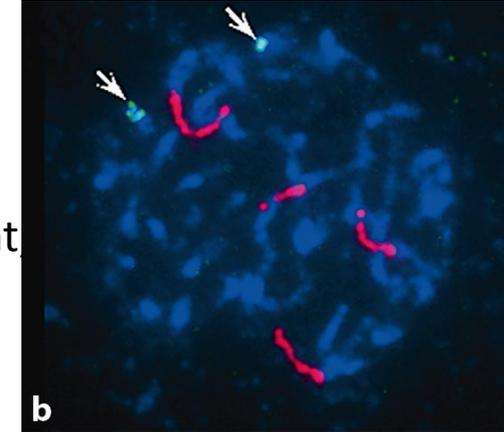
## 1.2) Quelques polymorphismes des génomes

- **Répétitions (duplication, tripllication... 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

- **Grandes répétitions et grands réarrangements chromosomique > La cytogénétique :**

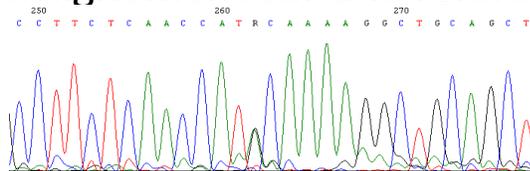
La coloration au Giemsa des chromosomes et l'étude des bandes claires et sombres qui apparaissent. L'hybridation *in situ* de sondes marquées (fluorescence) sur les chromosomes (**FISH**), sont des techniques très puissantes pour mettre en évidence ces évènements.



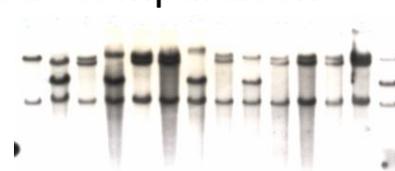
Détection de l'ADNr 18S-28S (rouge) et 5S (vert) par FISH sur des chromosomes de peuplier

- **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

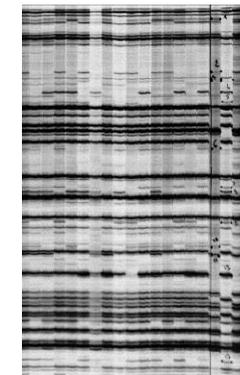


SNP



RFLP

*Restriction Fragment Length polymorphism*



AFLP  
*Amplified Fragment Length Polymorphism*

## 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 ATGGTAA**G**CCTGAC

Séquence ATGGTAA**A**CCTGAC

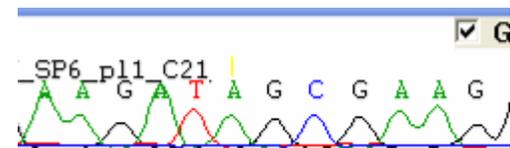


SNP



Sur un chromatogramme :

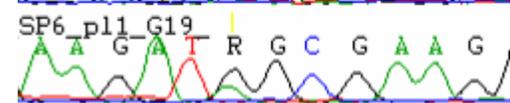
Homozygote 1



Homozygote 2



Hétérozygote



## 2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAA**G**CCTGAC

Séquence ATGGTAA**A**CCTGAC

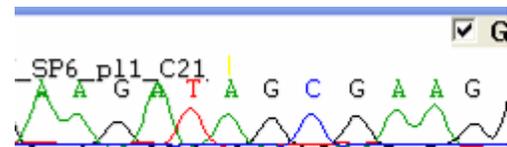


SNP



Sur un chromatogramme :

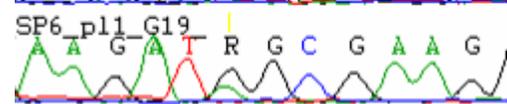
Homozygote 1



Homozygote 2



Hétérozygote



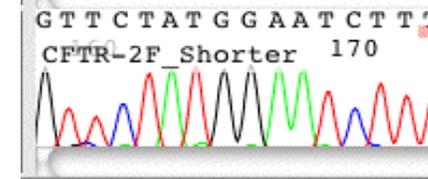
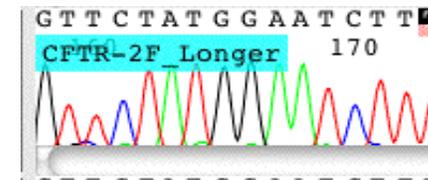
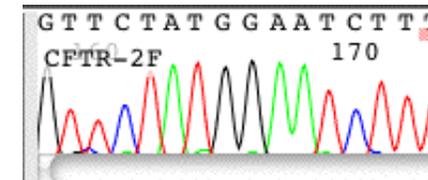
Référence 'TAGCGT-AT

Séquence 'TAGCGTCAT



indel

(Insertion/délétion)



## 2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAA**G**CCTGAC

Séquence ATGGTAA**A**CCTGAC

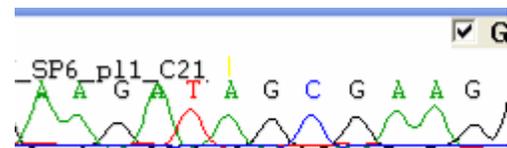


SNP



Sur un chromatogramme :

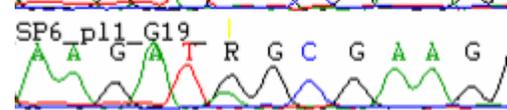
Homozygote 1



Homozygote 2



Hétérozygote



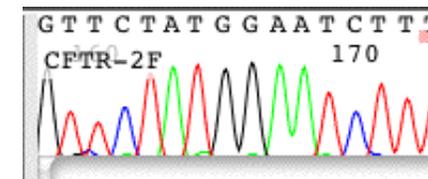
Référence 'TAGCGT-AT

Séquence 'TAGCGTCAT

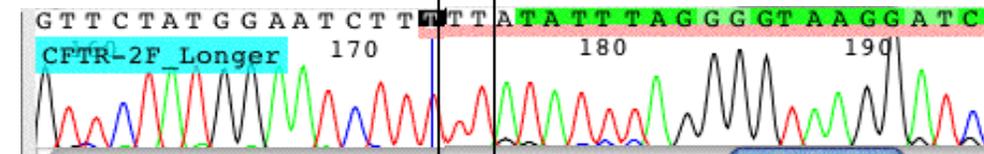


indel

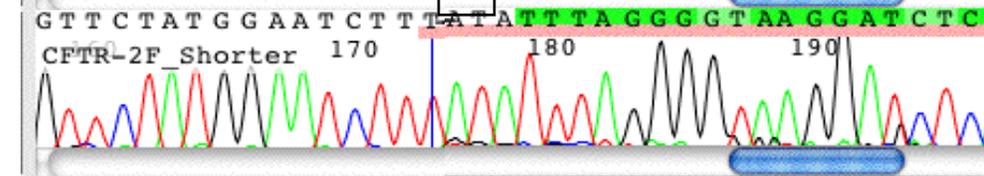
(Insertion/délétion)



Homozygote 1



Homozygote 2



## 2.1) Description des SNPs et Insertion/Délétion

Sur une séquence :

Référence ATGGTAA**G**CCTGAC

Séquence ATGGTAA**A**CCTGAC

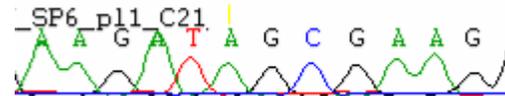


SNP

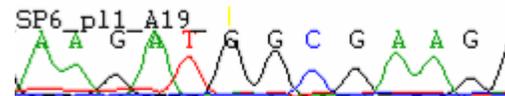


Sur un chromatogramme :

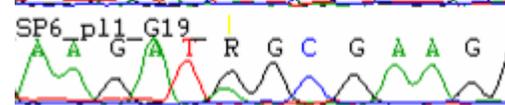
Homozygote 1



Homozygote 2



Hétérozygote



Référence 'TAGCGT-AT

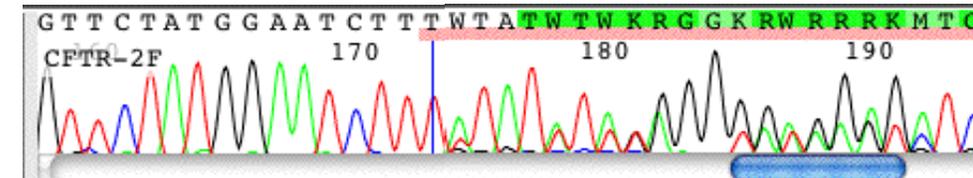
Séquence 'TAGCGTCAT



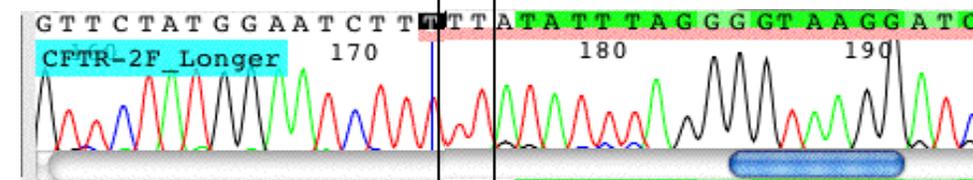
indel

(Insertion/délétion)

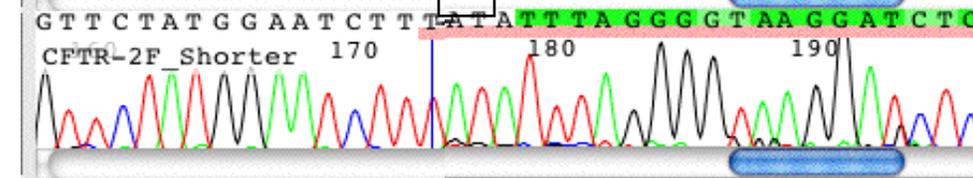
Hétérozygote



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

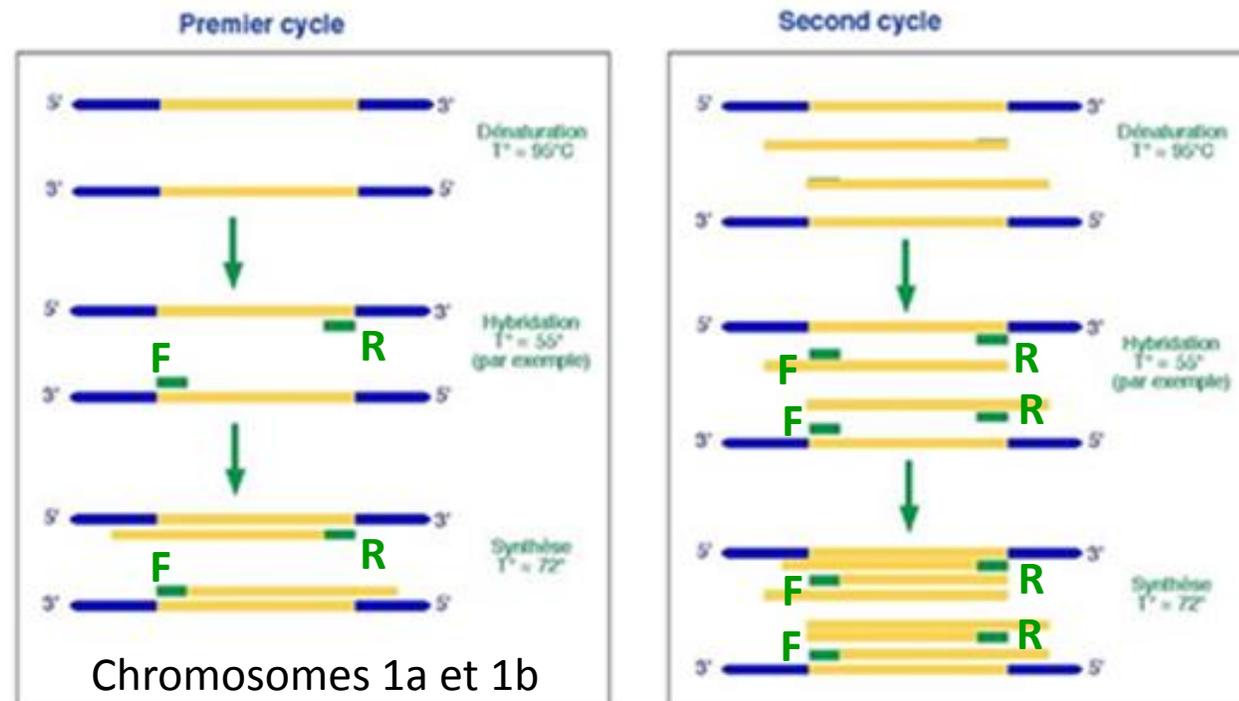
## 2.2) Description d'un outil d'étude des SNPs et InDel : Le séquençage de fragments PCR

- 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**.



F = amorce Forward

R = amorce Reverse

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

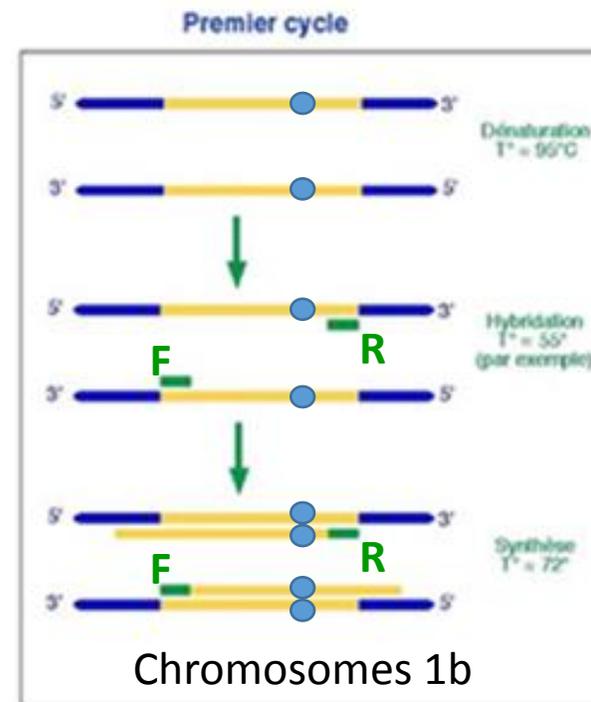
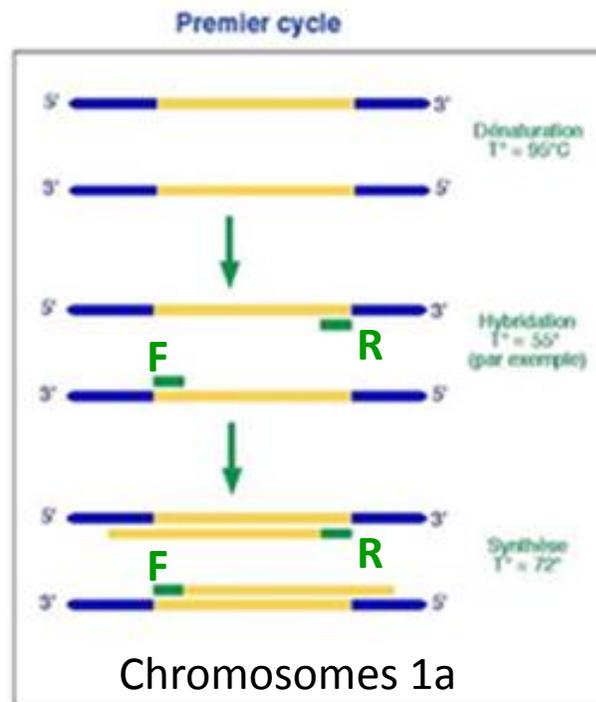
- La PCR sur génome diploïde hétérozygote

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



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 vite 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.

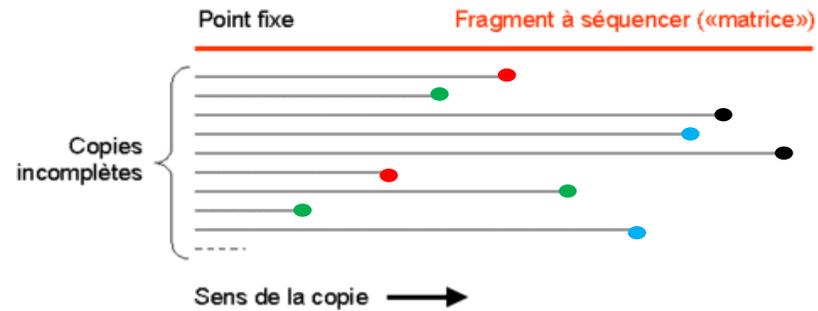
- **Le séquenceur**

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

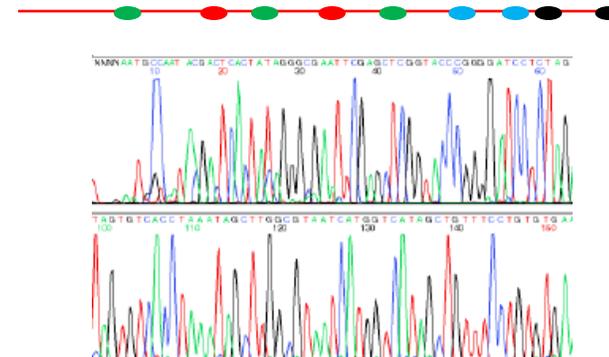
### 1) Synthèse



### 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 :

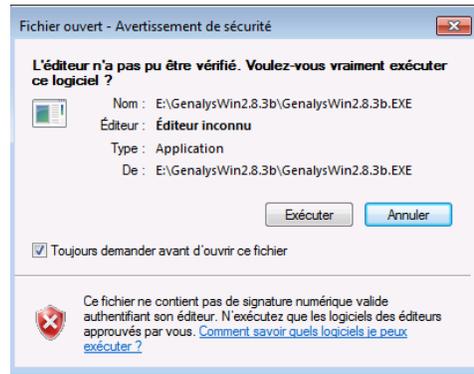
1<sup>ère</sup> ligne : > Nom\_seq Identifiant ; Organisme....

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.

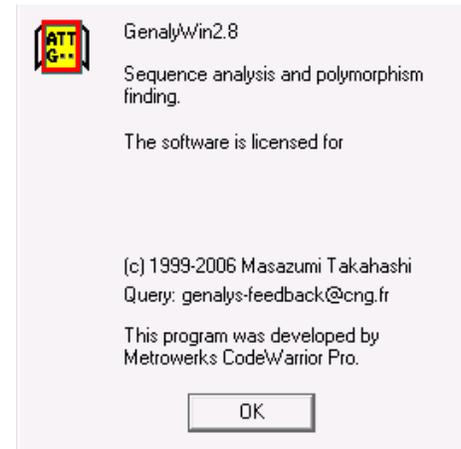
REMARQUE : En suivant ces règles, vous pouvez créer votre propre fichier de séquence.txt, Genalys pourra l'ouvrir !

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

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

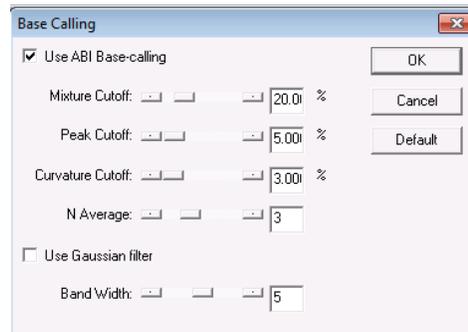


Exécuter



OK

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



OK

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

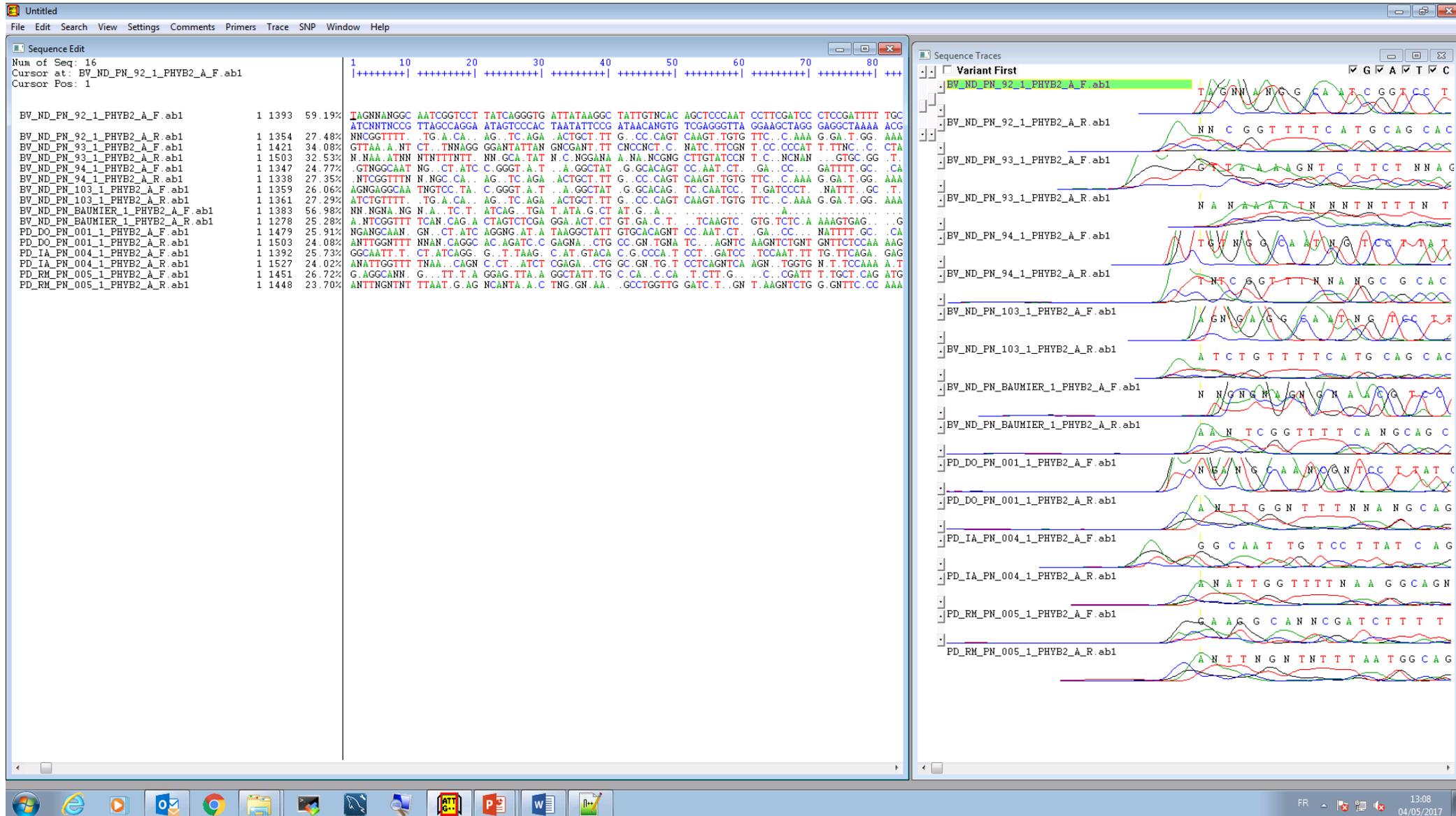
### 3.3) Ouvrir des fichiers de séquences (2/6)

- **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

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.

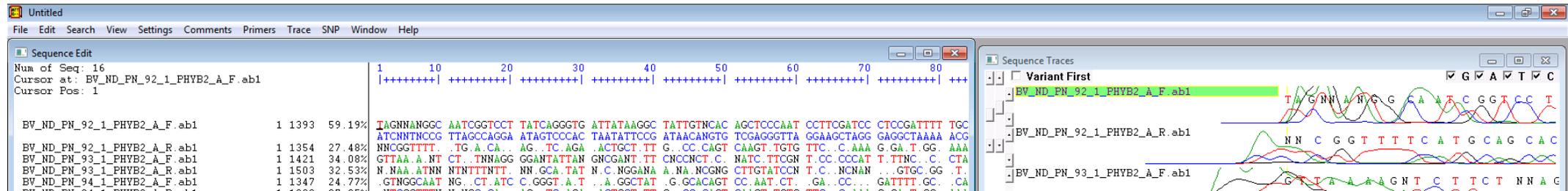
### 3.3) Ouvrir des fichiers de séquences (3/6)



The screenshot displays a software interface for sequence analysis, divided into two main panels. The left panel, titled "Sequence Edit", shows a list of 16 sequence files with their respective positions and coverage percentages. The right panel, titled "Sequence Traces", displays chromatograms for each sequence, with the first trace highlighted in green and labeled "Variant First".

File Name	Position	Coverage (%)	Sequence
BV_ND_PN_92_1_PHYB2_A_F.ab1	1 1393	59.19%	TAGNNANGCC AATCGGTCCT TATCAGGGTG ATTATAAGGC TATTGTNCAC AGCTOCCAAT CCTTCGATCC CTCGGATTTT TGC
BV_ND_PN_92_1_PHYB2_A_R.ab1	1 1354	27.48%	ATCNRNTCCG TTAGCCAGGA ATAGTCCAC TAATATTCG ATAACANGTG TOGAGGGTTA GSAAGCTAGG GAGGCTAAAA ACG
BV_ND_PN_93_1_PHYB2_A_F.ab1	1 1421	34.08%	NMCGGTTTT .TG.A.CA. AG .TC.AGA .ACTGCT.TT G .CC.CAGT CAAGT.TGTG TTC .C.AAA G.GA.T.GG. AAA
BV_ND_PN_93_1_PHYB2_A_R.ab1	1 1503	32.53%	GTTAA.A.NT CT .TNRWAGG GGANTATTAN GCGANT.TT CCCCCT.C. NATC.TTCGN T.CC.CCCAT T.TTWC .C. CTA
BV_ND_PN_94_1_PHYB2_A_F.ab1	1 1347	24.77%	N.MAA.ATNN NTNTTNTT NN.GCA.TAT N.C.NGANA A.NA.NCGNG CTGTATCCN T.C.NCHAM .GTGC.GG.T
BV_ND_PN_94_1_PHYB2_A_R.ab1	1 1347	24.77%	GTNGGCAAT NG .CT.ATC C.GGGT.A.T .A.GGCTAT G.GCACAGT CC.AAT.CT. GA.CC. .GATTTT.GC. .CA
BV_ND_PN_103_1_PHYB2_A_F.ab1	1 1338	27.35%	.NTCCGTTTT N.NGC.CA. AG .TC.AGA .ACTGCT.TT G .CC.CAGT CAAGT.TGTG TTC .C.AAA G.GA.T.GG. AAA
BV_ND_PN_103_1_PHYB2_A_R.ab1	1 1359	26.06%	ACNGAGGCAA TNGTCC.TA. C.GGGT.A.T .A.GGCTAT G.GCACAGT TC.CAATCC T.GATCCCT .NATTT.GC.T.
BV_ND_PN_103_1_PHYB2_A_F.ab1	1 1361	27.29%	ATCTGTTTT .TG.A.CA. AG .TC.AGA .ACTGCT.TT G .CC.CAGT CAAGT.TGTG TTC .C.AAA G.GA.T.GG. AAA
BV_ND_PN_BAUMIER_1_PHYB2_A_F.ab1	1 1383	56.98%	NN.NGNA.NG N.A..TC.T. ATCAG..TGA T.ATA.G.CT AT.G.A.. .TCAAGTC. GTG.TCTC.A.AAAGTGAG. .G
BV_ND_PN_BAUMIER_1_PHYB2_A_R.ab1	1 1278	25.28%	A.NTCGGTTT TCAN.CAG.A CTAGTCTCGA GGA.ACT.CT GT.GA.C.T. .TCAAGTC. GTG.TCTC.A.AAAGTGAG. .G
PD_DO_PN_001_1_PHYB2_A_F.ab1	1 1479	25.91%	NGANGCAAN. GN .CT.ATC AGNGG.AT.A TAAGGCTATT GTGCACAGNT CC.AAT.CT. .GA.CC. .NATTTT.GC. .CA
PD_DO_PN_001_1_PHYB2_A_R.ab1	1 1503	24.08%	ANTTGGNTT NNAN.CAGCC AC.AGATC.C GAGNA..CTG CC.GN.TGNA TC. .AGNTC.AAGNTCTGNT GNTTCTCCAA AAG
PD_IA_PN_004_1_PHYB2_A_F.ab1	1 1392	25.73%	GGCAATT.T. CT.ATCAGG. G..T.TAAG. C.AT.GTACA C.G.CCCA.T CCT..GATCC .TCCAAAT.TT TG.TTCAGAA.GAG
PD_IA_PN_004_1_PHYB2_A_R.ab1	1 1527	24.02%	ANATTGGTTT TNAA..CAGN G.CT..ATCT CGAGA..CTG GC.GN.TG.T CCTCAGNTCA AGN..TGGTG N.T.TCCAAA.A.T
PD_RM_PN_005_1_PHYB2_A_F.ab1	1 1451	26.72%	G.AGGCANN. G..TT.T.A GGAG.TTA.A GGCTATT.TG C.CA..C.CA .T.CTT.G. .C.CGATT T.TGCT.CAG.ATG
PD_RM_PN_005_1_PHYB2_A_R.ab1	1 1448	23.70%	ANTTNGTNT TTAAT.G.AG NCANTA.A.C TNG.GN.AA. .GCCTGGTTG GATC.T..GN T.AAGNTCTG G.GNTTC.CC.AAA

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



The screenshot shows a software window titled 'Untitled' with a menu bar (File, Edit, Search, View, Settings, Comments, Primers, Trace, SNP, Window, Help). The 'Sequence Edit' window displays a list of sequences and their alignment with a reference sequence. The 'Sequence Traces' window shows chromatograms for the same sequences, with a 'Variant First' section highlighting a specific variant.

Sequence Name	Start	End	Identity %	Sequence
BV_ND_PN_92_1_PHYB2_A_F.ab1	1	1393	59.19%	TAGNNANGGC AATCGGTCCT TATCAGGGTG ATTATAAGGC TATTGTNCAC AGCTCCCAAT CCTTCGATCC CTCGATTTT TGC
BV_ND_PN_92_1_PHYB2_A_R.ab1	1	1354	27.48%	ATCNNNTCCG TTAGCCAGGA ATAGTCCAC TAATATTCGG ATAACANGTG TCGAGGGTTA GGAAGCTAGG GAGGCTAAAA ACG
BV_ND_PN_93_1_PHYB2_A_F.ab1	1	1421	34.08%	NNCGGTTTT .TG.A.CA. .AG.TC.AGA .ACTGCT.TT G..CC.CAGT CAAGT.TGTG TTC..C.AAA G.GA.T.GG. AAA
BV_ND_PN_93_1_PHYB2_A_R.ab1	1	1503	32.53%	GTTAA.A.NT.CT.TNNAGG GGANTATTAN GNCGANT.TT CNOCNCT.C. NATC.TTCGN T.CC.CCCAT T.TTNC..C. CTA
BV_ND_PN_94_1_PHYB2_A_F.ab1	1	1347	24.77%	N.NAA.ATNN NTNTTNTT. NN.GCA.TAT N.C.NGGANA A.NA.NCGNG CTTGTATCCN T.C..NCNAN ..GTGC.GG .T.

- **Deux fenêtres se sont ouvertes :**

- Sequence Edit : séquence(s) lue(s) à partir des fichier.ab1
- Sequence Traces : chromatogramme(s) correspondant

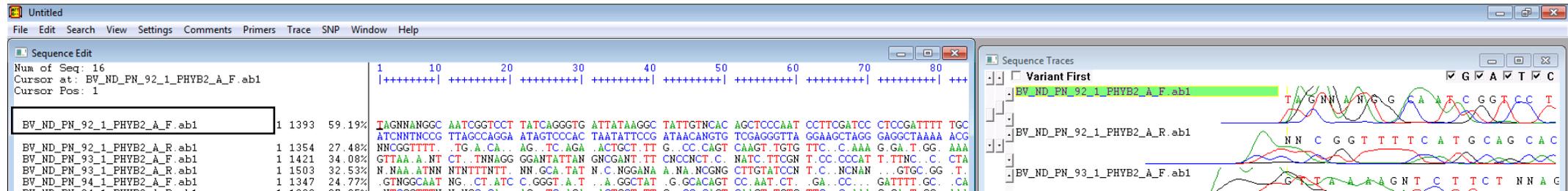
En ouvrant un fichier.txt, seule la fenêtre « Sequence Edit » sera informée

- **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 1<sup>ère</sup> 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)



The screenshot shows a software window titled 'Untitled' with a menu bar (File, Edit, Search, View, Settings, Comments, Primers, Trace, SNP, Window, Help). The main area is split into two panes. The left pane, 'Sequence Edit', displays a list of sequences on the left and a sequence alignment on the right. The right pane, 'Sequence Traces', shows a chromatogram of the selected sequence.

Sequence Name	Position	Percentage	Sequence
BV_ND_PN_92_1_PHYB2_A_F.ab1	1	59.19%	TAGNNANGGC AATCGGTCCT TATCAGGGTG ATTATAAGGC TATTGTNCAC AGCTCCCAAT CCTTCGATCC CTCGATTTT TGC
BV_ND_PN_92_1_PHYB2_A_R.ab1	1	27.48%	ATCNNNCCG TTAGCCAGGA ATAGTCCAC TAATATTCG ATAACANGTG TCGAGGGTTA GGAAGCTAGG GAGGCTAAAA ACG
BV_ND_PN_93_1_PHYB2_A_F.ab1	1	34.08%	NACGGTTTT .TG.A.CA. .AG.TC.AGA .ACTGCT.TT G..CC.CAGT CAAGT.TGTG TTC..C.AAA G.GA.T.GG. AAA
BV_ND_PN_93_1_PHYB2_A_R.ab1	1	32.53%	GTTAA.A.NT.CT.TNNAGG GGANTATTAN GNCGANT.TT CNOCNCT.C NATC.TTCGN T.CC.CCCAT T.TTNC.C.CTA
BV_ND_PN_94_1_PHYB2_A_F.ab1	1	24.77%	N.NAA.ATNN NTNTTNTT. NN.GCA.TAT N.C.NGGANA A.NA.NCGNG CTTGTATCCN T.C..NCNAN ..GTGC.GG.T.

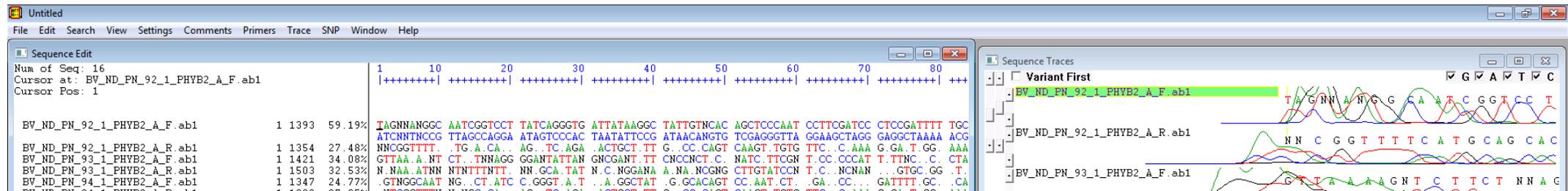
- **Sequence Edit** , elle est composée de 2 parties :

A droite

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

- 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)



The screenshot shows the Genalys software interface. The 'Sequence Edit' window on the left displays a list of sequence files and their statistics:

File Name	Num of Seq	Cursor at	Cursor Pos	Sequence
BV_ND_PN_92_1_PHYB2_A_F.ab1	1	1393	59.19%	TAGNNANGGC AATCGGTCCT TATCAGGGTG ATTATAAGGC TATTGTNCAC AGCTCCCAAT CCTTCGATCC CTCGATTTT TGC
BV_ND_PN_92_1_PHYB2_A_R.ab1	1	1354	27.48%	ATCNTNCCG TTAGCCAGGA ATAGTCCAC TAATATTCG ATAACANGTG TCGAGGGTTA GGAAGCTAGG GAGGCTAAAA ACG
BV_ND_PN_93_1_PHYB2_A_F.ab1	1	1421	34.08%	NACGGTTTT .TG.A.CA. .AG.TC.AGA .ACTGCT.TT G..CC.CAGT CAAGT.TGTG TTC..C.AAA G.GA.T.GG. AAA
BV_ND_PN_93_1_PHYB2_A_R.ab1	1	1503	32.53%	GTTAA.A.NT.CT..TNNAGG GGANTATTAN GNCGANT.TT CNOCNCT.C. NATC.TTCGN T.CC.CCCAT T.TTNC..C.CTA
BV_ND_PN_94_1_PHYB2_A_F.ab1	1	1347	24.77%	N.NAA.ATNN NTNTTNTT. NN.GCA.TAT N.C.NGGANA A.NA.NCGNG CTTGTATCCN T.C..NCNAN ..GTGC.GG..T.

The 'Sequence Traces' window on the right shows a chromatogram with four tracks corresponding to the files listed in the 'Sequence Edit' window. The tracks are labeled 'Variant First', 'BV\_ND\_PN\_92\_1\_PHYB2\_A\_F.ab1', 'BV\_ND\_PN\_92\_1\_PHYB2\_A\_R.ab1', and 'BV\_ND\_PN\_93\_1\_PHYB2\_A\_F.ab1'. The chromatogram displays peaks for the four bases (A, T, G, C) across the sequence length.

- **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 disparaître les pics correspondant à l'une des quatre bases A, T, G, C en décochant les boîtes 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 reverse-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 séquences Reverse : en vert (elles sont reverse-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.

### • 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%

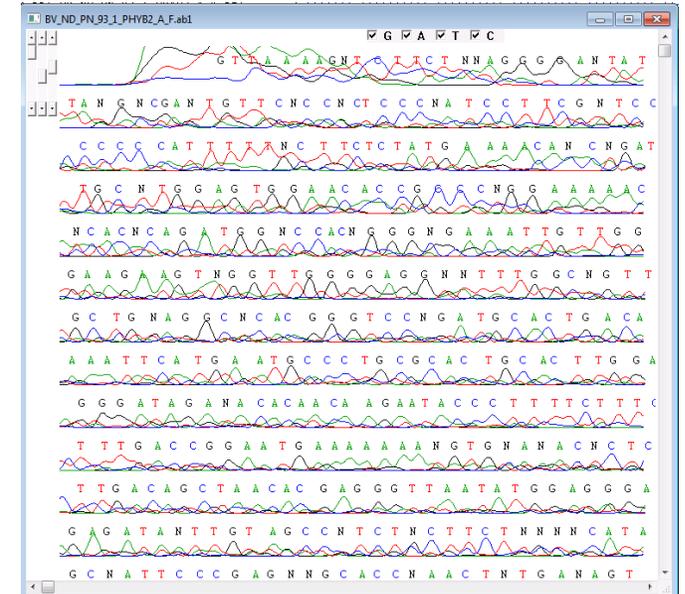
⇒ Suppression de la séquence (cf diapo 21)

⇒ 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
```



Sequence Name	Start	End	Identity (%)
PHYB2_AMPLICON_POPULUS_NIGRA	1	663	100.00%
BV_ND_PN_92_1_PHYB2_A_F.ab1	31	1393	87.99%
BV_ND_PN_92_1_PHYB2_A_R.ab1	24	1354	96.00%
BV_ND_PN_94_1_PHYB2_A_F.ab1	36	1347	97.93%
BV_ND_PN_94_1_PHYB2_A_R.ab1	08	1338	95.81%
BV_ND_PN_103_1_PHYB2_A_F.ab1	35	1359	94.02%

=> 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

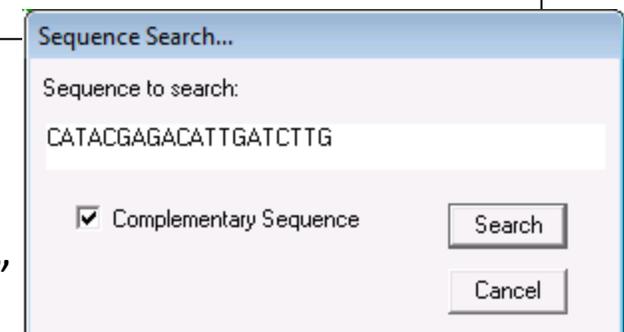
- Ouvrir le fichier amorces\_PHYB2.txt.
- Copier la sequence PHYB2\_F\_test
- Dans la fenêtre "Sequence Edit", mettre le curseur sur le nom de la reference
- Search/Sequence Search...
- Coller la sequence PHYB2\_F et « Search »
- Clic droit au-dessus de la séquence surlignée et ajouter un commentaire.

```
>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, sélectionner **Complementary Sequence** et "Search"
- Clic droit au-dessus de la séquence surlignée et ajouter un commentaire.



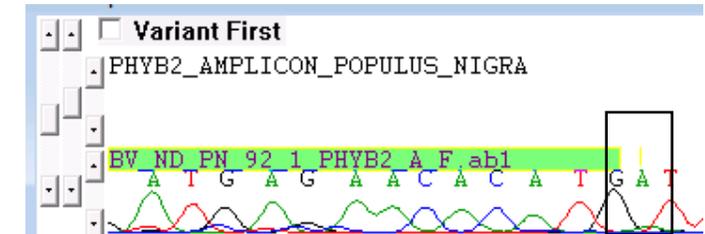
## 3.4) Aligner des séquences et commencer l'analyse

### • Analyse – Suppression de séquences inutiles

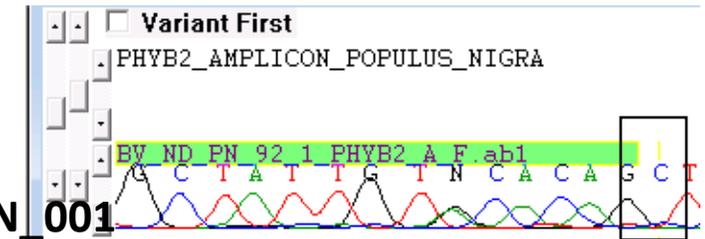
- Sélectionner la séquence en amont de la position 1 et supprimer (backsupp)
- Sélectionner la séquence en aval de la position 663 et supprimer (backsupp)

### • Analyse – Nettoyage et amélioration de l'alignement

- **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**



- **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**



- **Pos\_76** : clic droit sur la séquence de **BV\_ND\_PN\_BAUMIER**
  - Supprimer la base en trop.

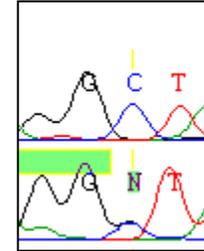
- **Regarder** les chromatogrammes en amont de la position 70
  - **Sélectionner** (en se mettant à l'extrémité gauche de la fenêtre **et supprimer** les séquences en amont de la position 70 (backsupp)

## 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

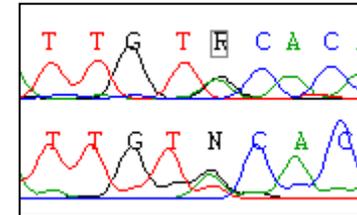


6

- **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**

- **Pos\_106** : clic droit sur la séquence de **BV\_ND\_PN\_92\_1\_F** et **\_R** : R

10 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

- **Pos\_79** : sélectionner la base de la référence où il y a le SNP

Clic droit au-dessus de cette base

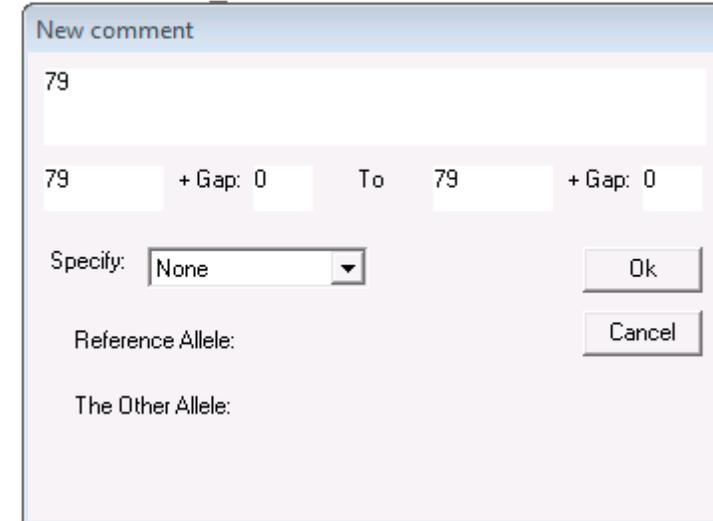
Une fenêtre « New comment » s'ouvre.

17

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**

- **Pos\_79** : sélectionner la base de la référence où il y a le SNP

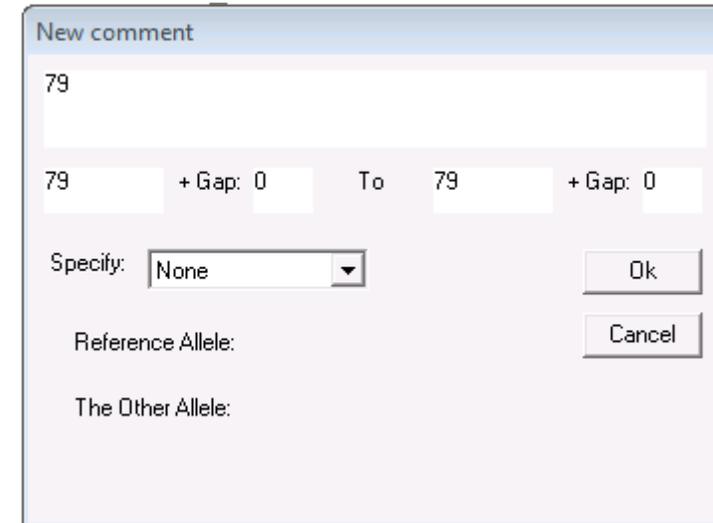
Clic droit au-dessus de cette base

Une fenêtre « New comment » s'ouvre.

17 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 :

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

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

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 branches: 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**

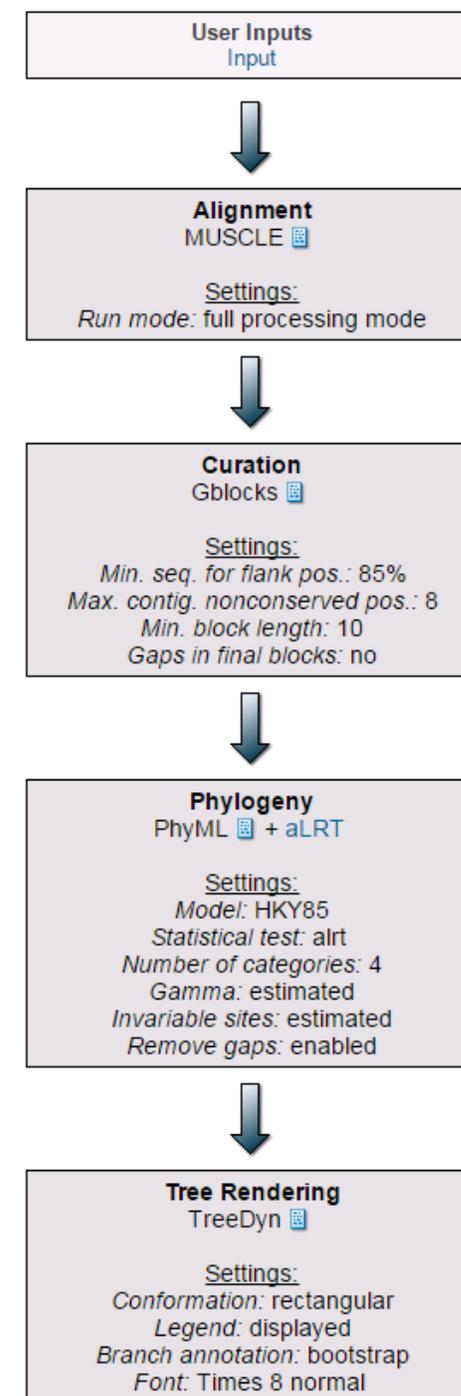
### Process Info:

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

### Computation:

Overall time: 0 seconds

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



## 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

### Tree Rendering results

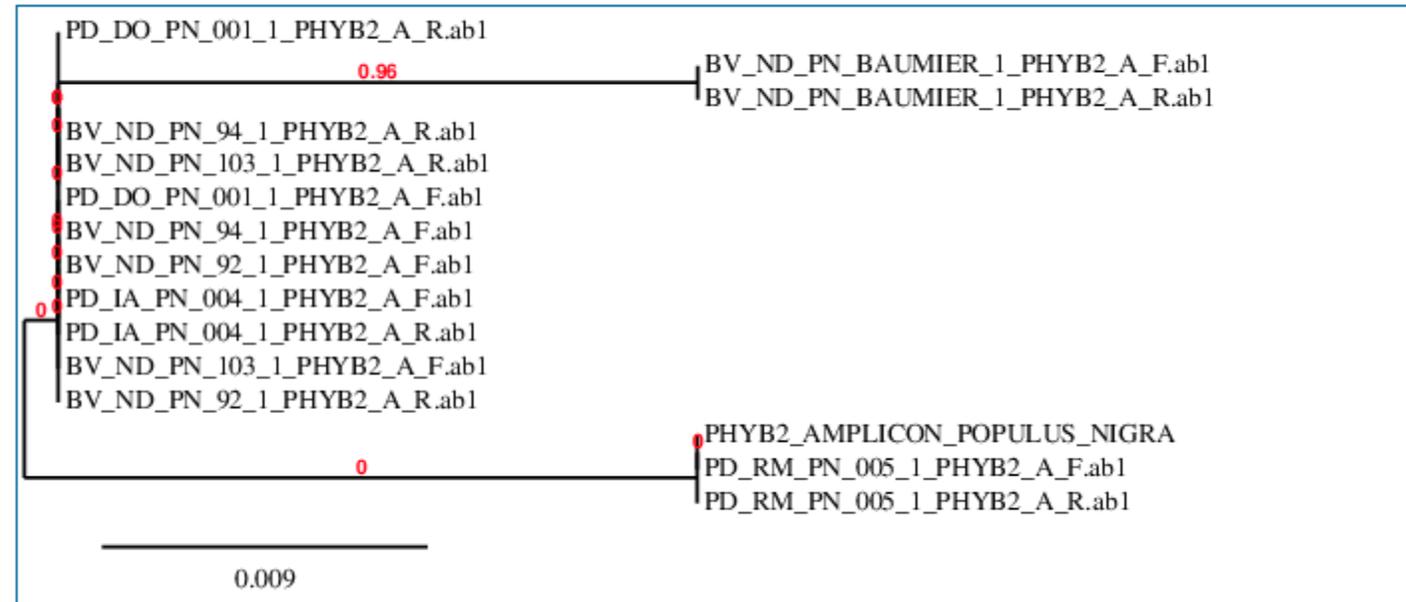


Figure 1: Phylogenetic tree.

## 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

Select an action and click leaf or internal branch:

Colorize  leaf  branch choose a color   and a legend label

Reroot (outgroup)

Flip (flip an entire tree at a node)

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

### Tree Rendering results

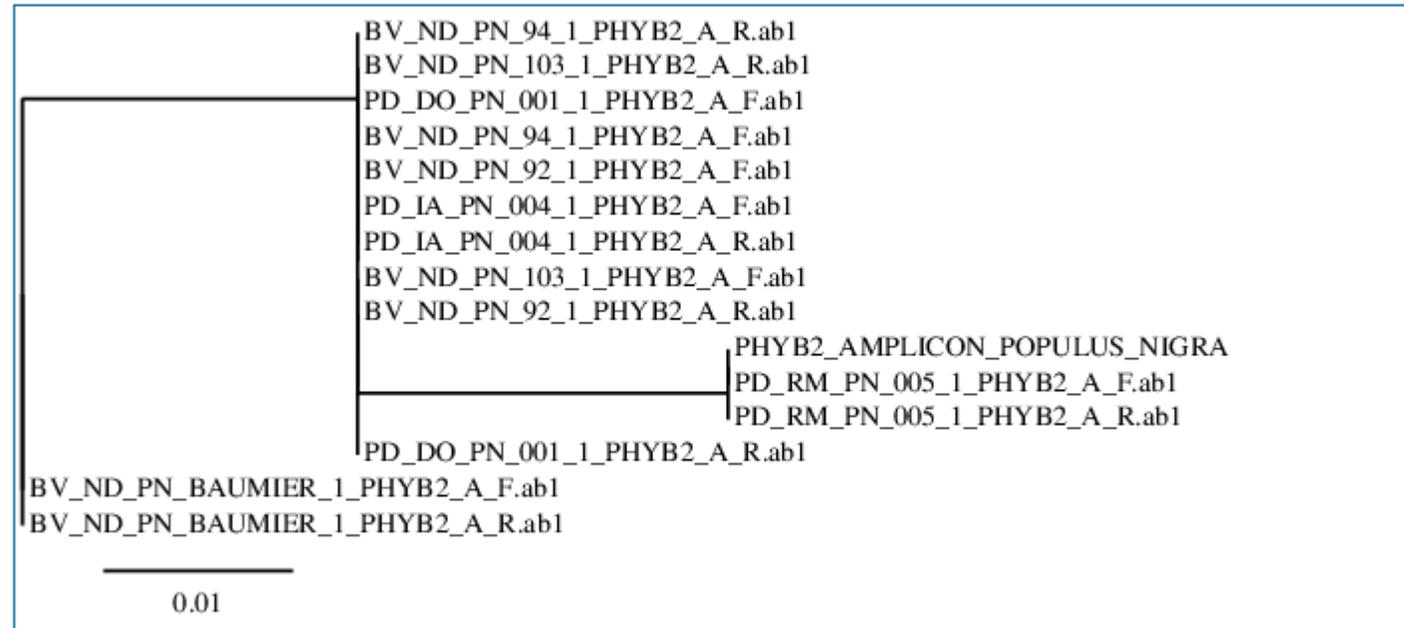


Figure 1: Phylogenetic tree.

## 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

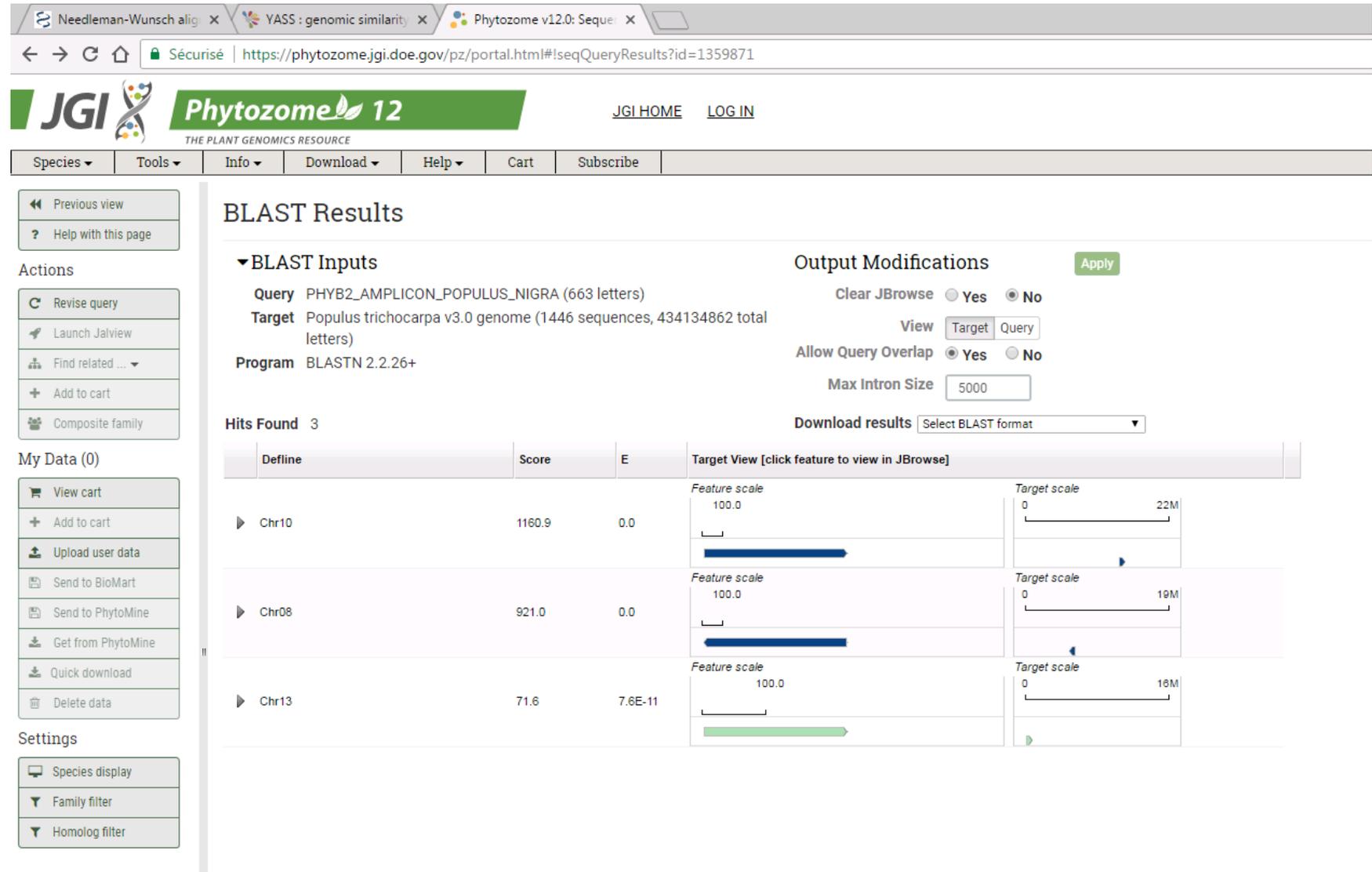
- 2.Build your query / Coller la séquence de la référence >PHYB2\_AMPLICON\_POPULUS\_NIGRA et

GO

## 4) Exploiter les SNP

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

- Cliquer sur le meilleur résultat



The screenshot shows the Phytozome v12.0 BLAST Results page. The browser tabs include Needleman-Wunsch align, YASS : genomic similarity, and Phytozome v12.0: Sequ. The URL is https://phytozome.jgi.doe.gov/pz/portal.html#!seqQueryResults?id=1359871. The page header features the JGI Phytozome 12 logo and navigation links for JGI HOME and LOG IN. A top navigation bar contains Species, Tools, Info, Download, Help, Cart, and Subscribe. The main content area is titled 'BLAST Results' and includes sections for BLAST Inputs, Output Modifications, Hits Found, and a table of results. The BLAST Inputs section shows a query of PHYB2\_AMPLICON\_POPULUS\_NIGRA (663 letters) and a target of Populus trichocarpa v3.0 genome (1446 sequences, 434134862 total letters). The Output Modifications section includes options for Clear JBrowse, View (Target/Query), Allow Query Overlap, and Max Intron Size (5000). The Hits Found section shows 3 results, with a table listing the chromosome, score, E-value, and target view for each hit.

**BLAST Results**

▼BLAST Inputs

**Query** PHYB2\_AMPLICON\_POPULUS\_NIGRA (663 letters)  
**Target** Populus trichocarpa v3.0 genome (1446 sequences, 434134862 total letters)  
**Program** BLASTN 2.2.26+

**Output Modifications** Apply

Clear JBrowse  Yes  No  
 View    
 Allow Query Overlap  Yes  No  
 Max Intron Size

**Download results**

**Hits Found** 3

Define	Score	E	Target View [click feature to view in JBrowse]
▶ Chr10	1160.9	0.0	Feature scale 100.0 Target scale 0 22M
▶ Chr08	921.0	0.0	Feature scale 100.0 Target scale 0 19M
▶ Chr13	71.6	7.6E-11	Feature scale 100.0 Target scale 0 16M

**Actions**

- ◀ Previous view
- ? Help with this page
- Revise query
- Launch Jalview
- Find related ...
- Add to cart
- Composite family

**My Data (0)**

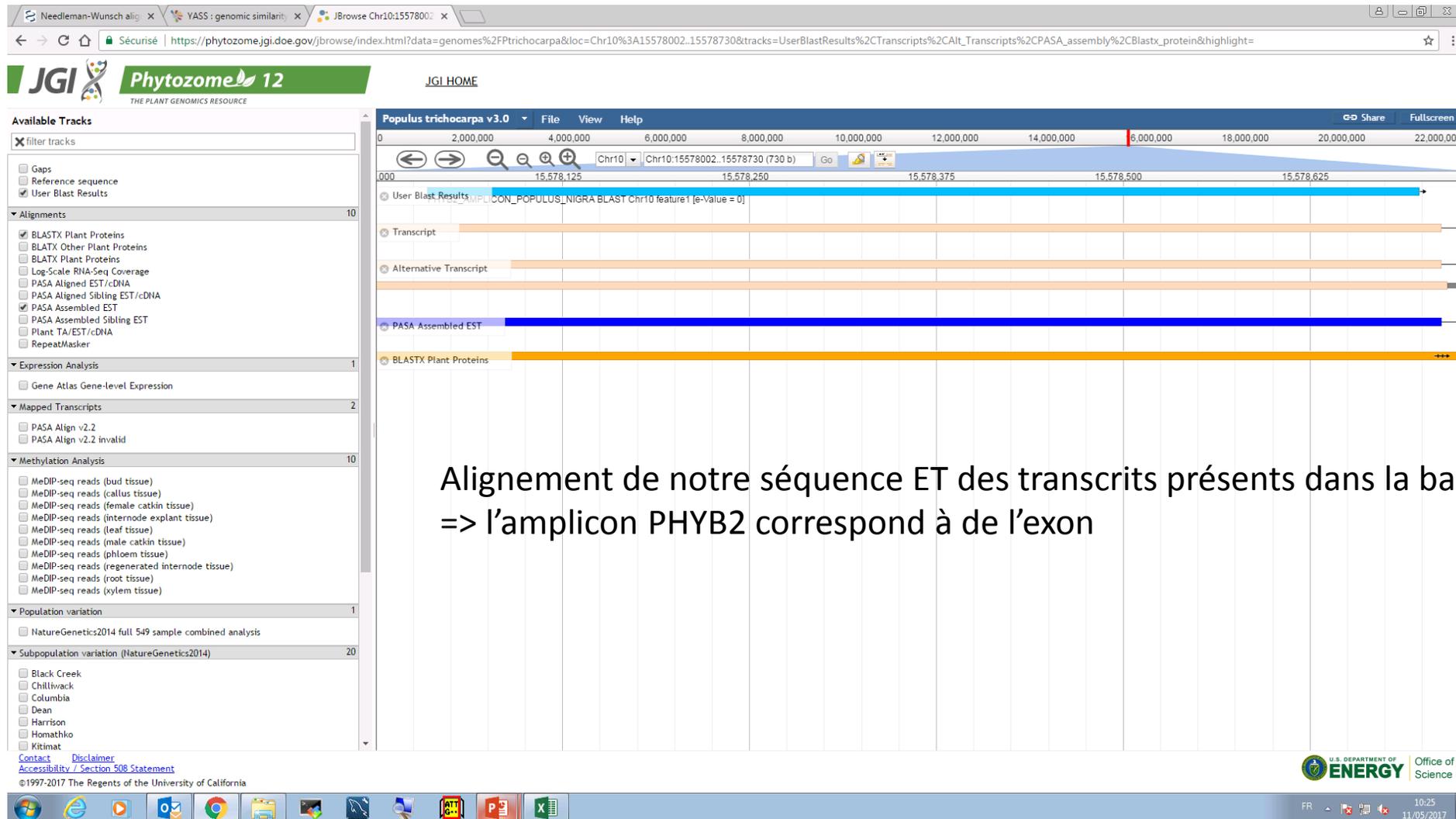
- View cart
- Add to cart
- Upload user data
- Send to BioMart
- Send to PhytoMine
- Get from PhytoMine
- Quick download
- Delete data

**Settings**

- Species display
- Family filter
- Homolog filter

## 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

```

TGTGGTCAAGGATGTTACAGGTCAAAAAGTGGTAATGGACAAATACGTCTTATACAAGGTGATTATAAGGCT
ATTGTGCACAGTCCCAATCCTTCGATCCCTCCGATTTTTGCTTCAGATGAGAACACATGTTGCTTGGAGTGGAA
CACTGCCATGSAAAAACTCACAGGATGGTCCAGGGGGGGAAGTTGTTGGGAAGATGTTGGTTGGGGAGSTTTT
TGGCAGTTGCTGTAGGCTCAAGGGTCCAGATGCACTGACAAAATTCATGATTGDCCTGCACAATGCAATTGGA
GGGATAGATACAGACAAGTTACCTTTTCATTCTTTGACCGGAATGAAAAAATGTGCAAACTCTCTTGACAG
CTAACCAAGAGGGTTAATATGGAGGGAGATATTATTGGAGCCTTCTGCTTCTTGACAGATTGCAATTCCTGAGTT
GCAGCAAACTTTGAAAGTTCAGAAACAGCAGGAAAAAATACTTTGCAAGGATGAAAGAGTTGGCTTACAT
TTGCCAGGAAATAAAAAATCCTTTGAGTGGTATACGCTTTACCAACTCACTTTTGGAGAACACAGACTTGACTG
AGGATCAACAGCAGTTTCTCGAGACTAGTGTGCTGATGTGAAAAACAGATATTGAAGATCATAOGAGACATTG
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 (1<sup>ère</sup> méthionine)
- **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 !