



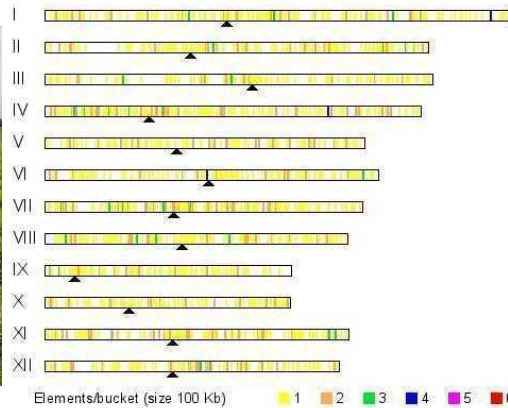
Basic notions In Annotation

Example of Transposable Elements

Le but du TP est de vous donner une idée rapide des méthodologies d'annotation des éléments transposables dans une grande région génomique, grâce à l'utilisation d'outils visuel, Gepard et Artemis.

De plus, vous serez amené à manipuler Artemis, ce qui vous aidera fortement pour le prochain TD

Cultivated Rice, *Oryza sativa* ssp *Japonica*



Facts & Data

Genome Size ~400Mb

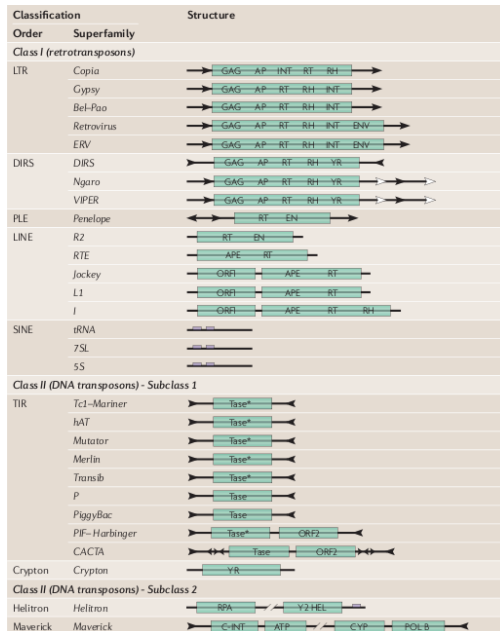
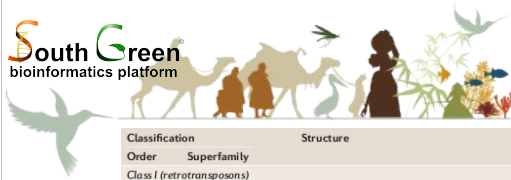
~ 25'000 'genes'

Repeats size 100bp → 14'000bp

Repeat frequency 2 → 3'000 times

Nous allons travailler sur le riz Asiatique, qui a un petit génome, et peu d'éléments transposables, comparativement à des plantes comme l'orge ou le blé.

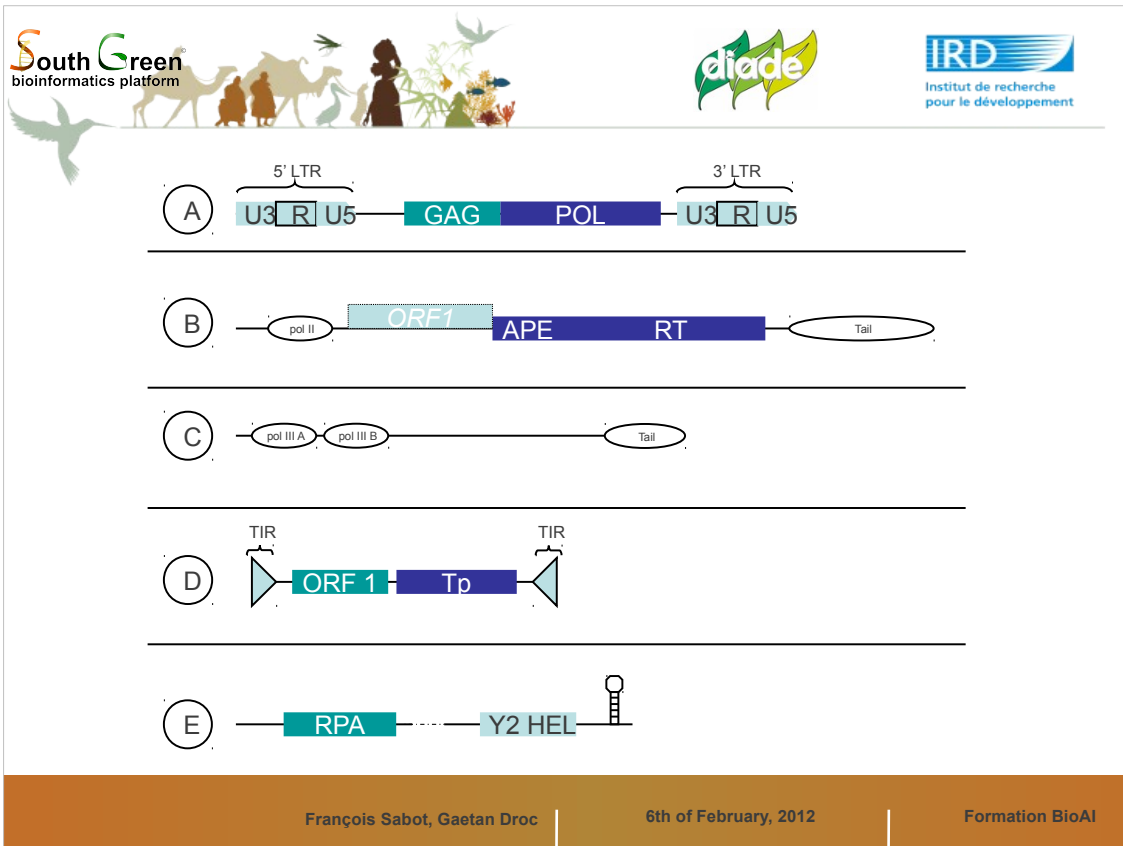
Paradoxalement, il est parfois plus dur d'annoter un génome avec peu d'éléments.



Wicker et al, 2007

Nature Reviews Genetics

La classification des éléments transposables, ou TE, est un sujet complexe. Une bonne idée de leur diversité est présentée dans le papier cité ci-dessus

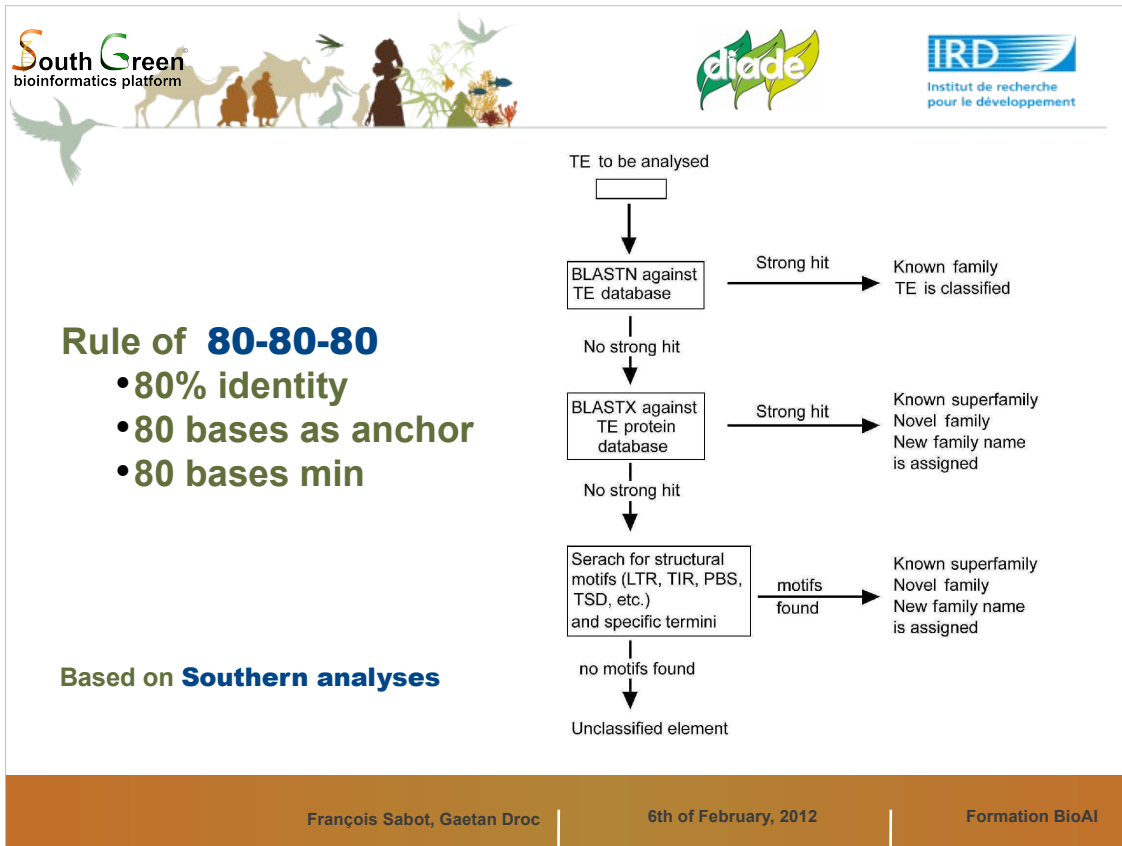


Une classification simplifiée des éléments présents dans les plantes.

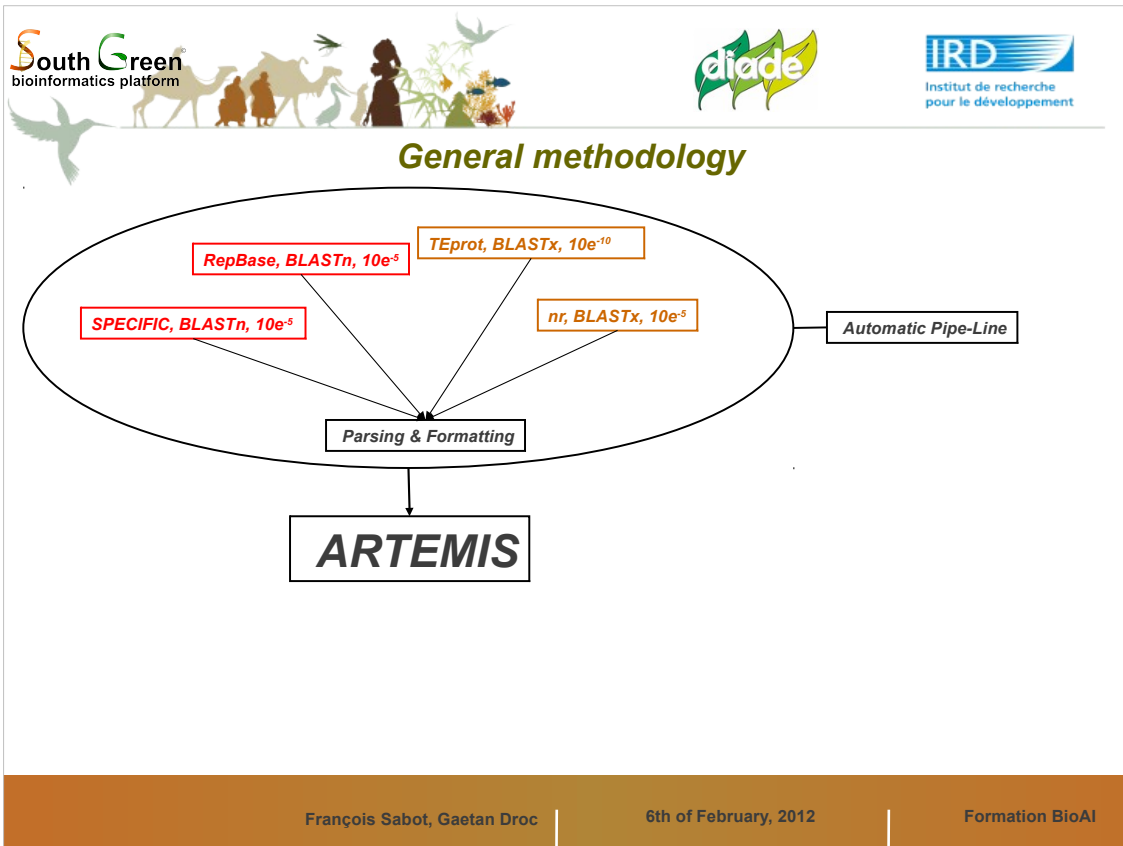
LTR = Long Terminal Repeats

TIR = Terminal Inverted Repeats

Les éléments sont en général bordés par des TSD, target site duplication, obtenus via une coupure asymétrique du double brin d'ADN au site d'insertion.



Ici un protocole basique d'annotation de TEs. La regles des 80 – 80 – 80 marche a peu pres bien en automatique, et évite de créer des faux positifs. Par contre elle ne permet pas de récupérer toutes les répétitions...



Voici un type de pipeline d'annotation automatique, qui fournit un set de données à utiliser ensuite dans Artemis pour finaliser le travail



Comparison Tools

- BLAST + Specific DB
- RepeatMasker
- Censor
- PlotRep
- Dotter/Gepard





Différents types d'outils existent. Ceux de comparaisons réclament une base de données de séquences pré-existantes



Prediction/Integrated **Tools**

- LTR_STRUC / LTRharvest / LTR-finder
- RepeatScout / RepeatFinder
- REPET

Certains outils permettent de prédire *ab initio*, donc sans connaissance préalable de séquences, simplement via des structures spécifiques des éléments transposables.



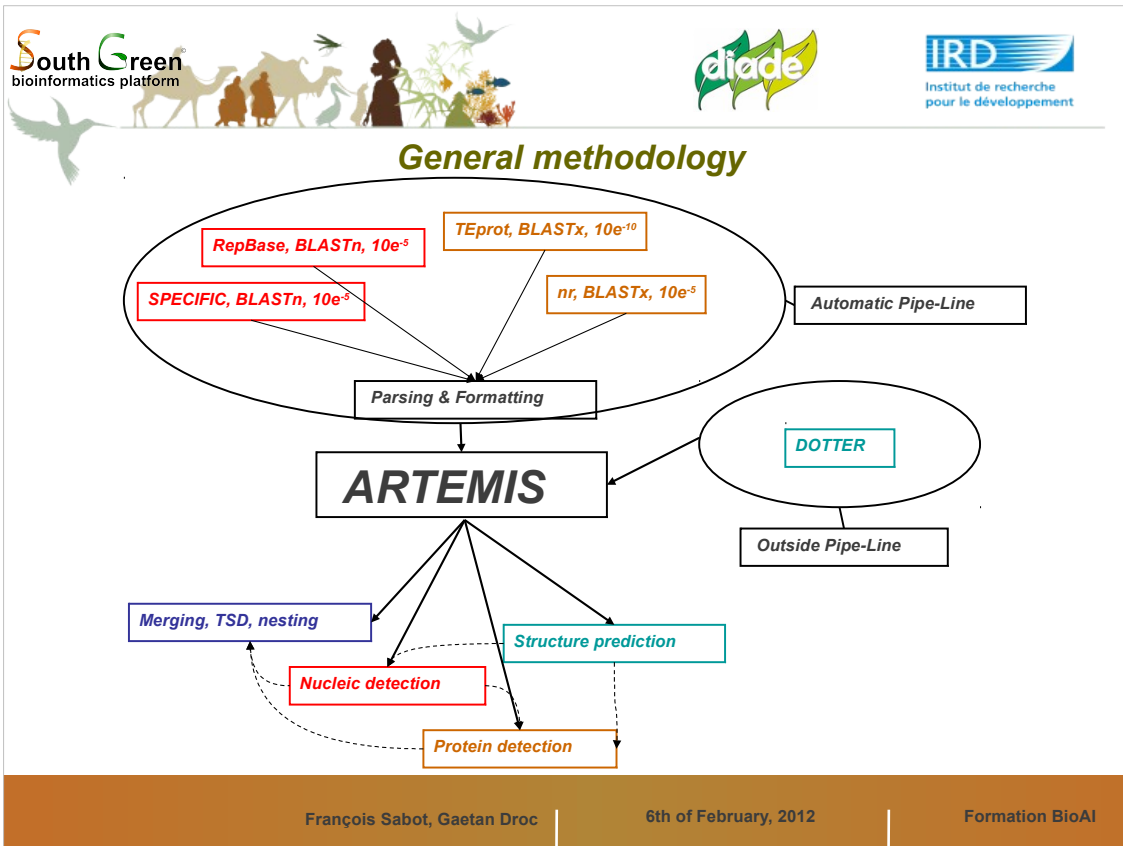
Automatic prediction only for **LTR Retrotransposons**

François Sabot, Gaetan Droc

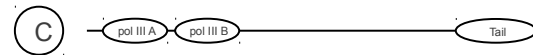
6th of February, 2012

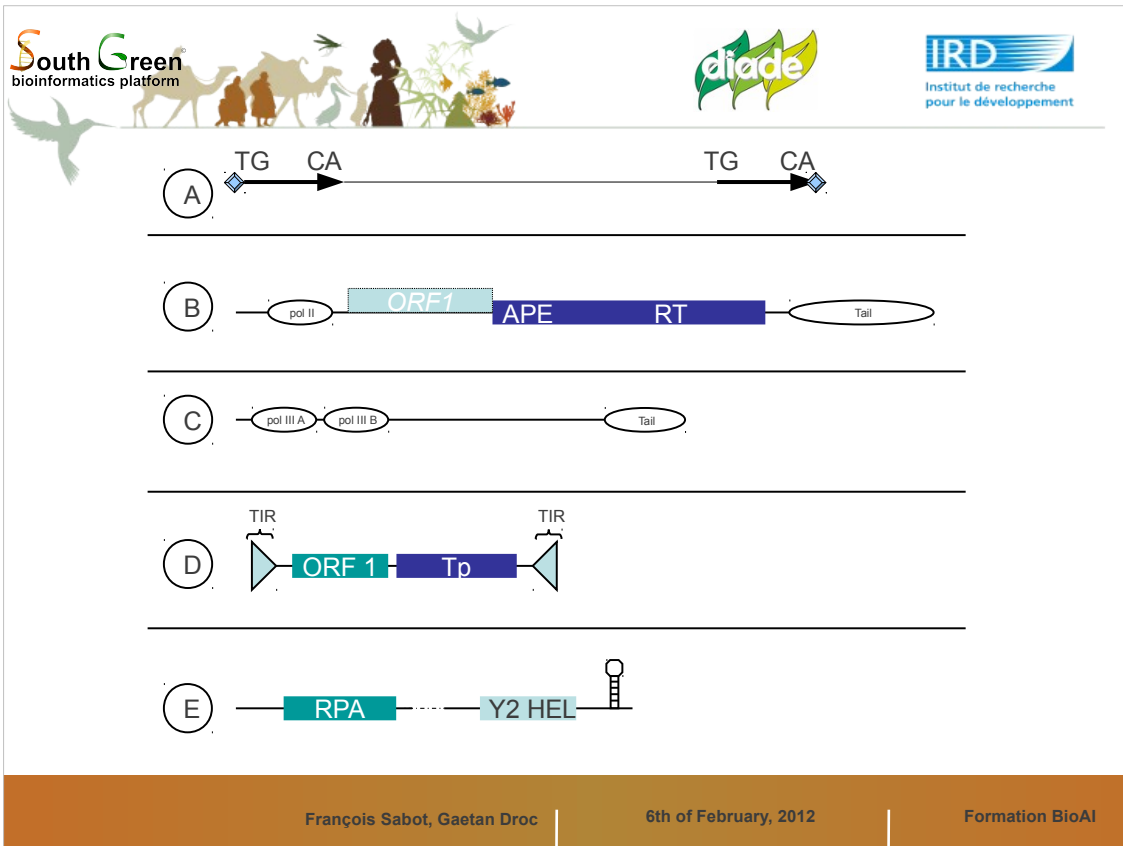
Formation BioAI

Malheureusement, la prédiction *ab initio* ne marche que pour les LTR rétrotransposons.



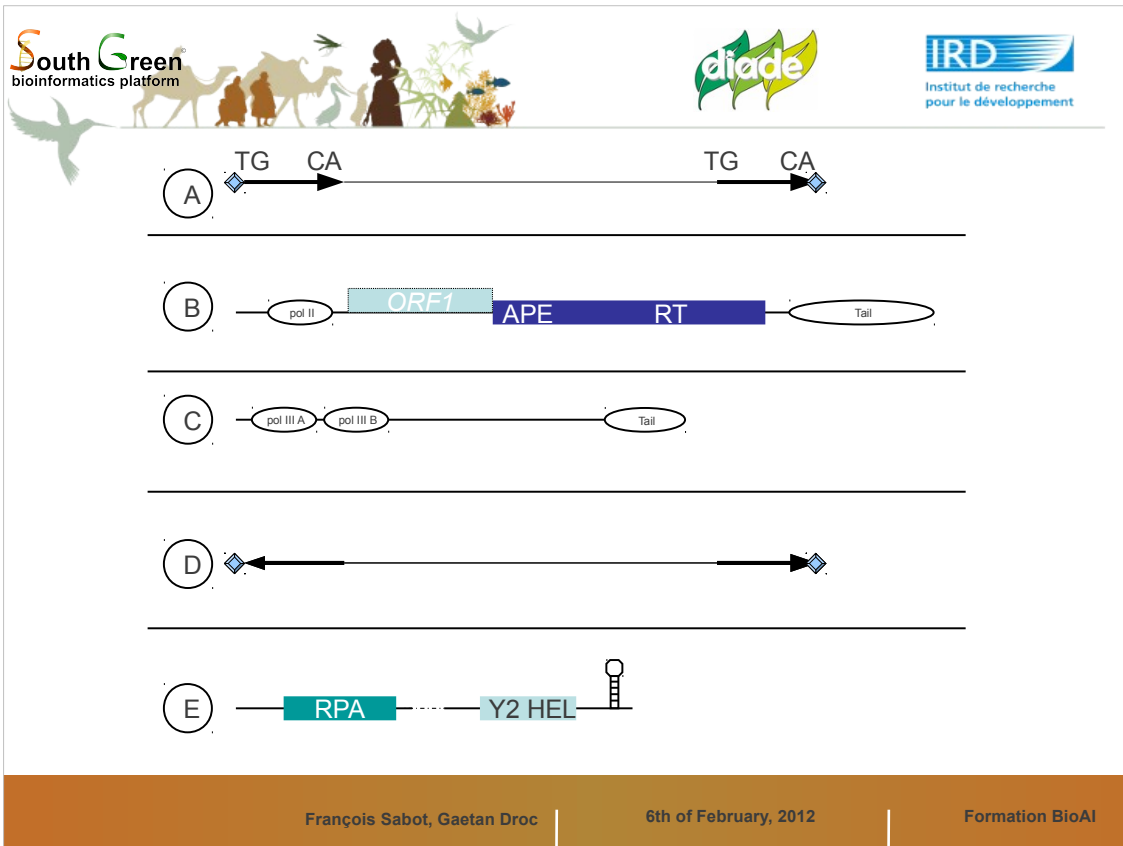
En plus du pipeline automatique, nous utiliserons des outils de type Dotter, puis grâce à Artemis nous allons identifier et annoter des structures d'éléments transposables.



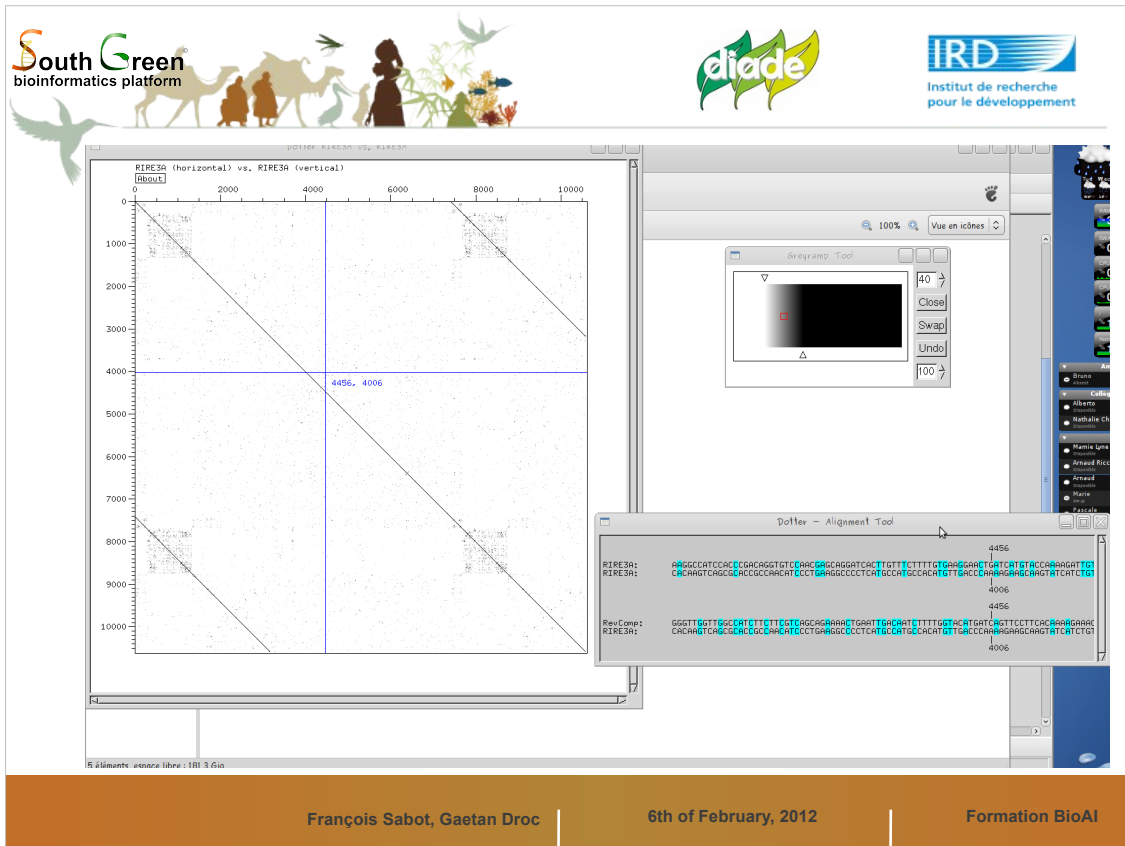


Les LTR commencent par TG et finissent par CA dans 99% des cas. Ils sont identiques entre eux au moment de l'insertions.

Les deux TSD sont de 5 bases

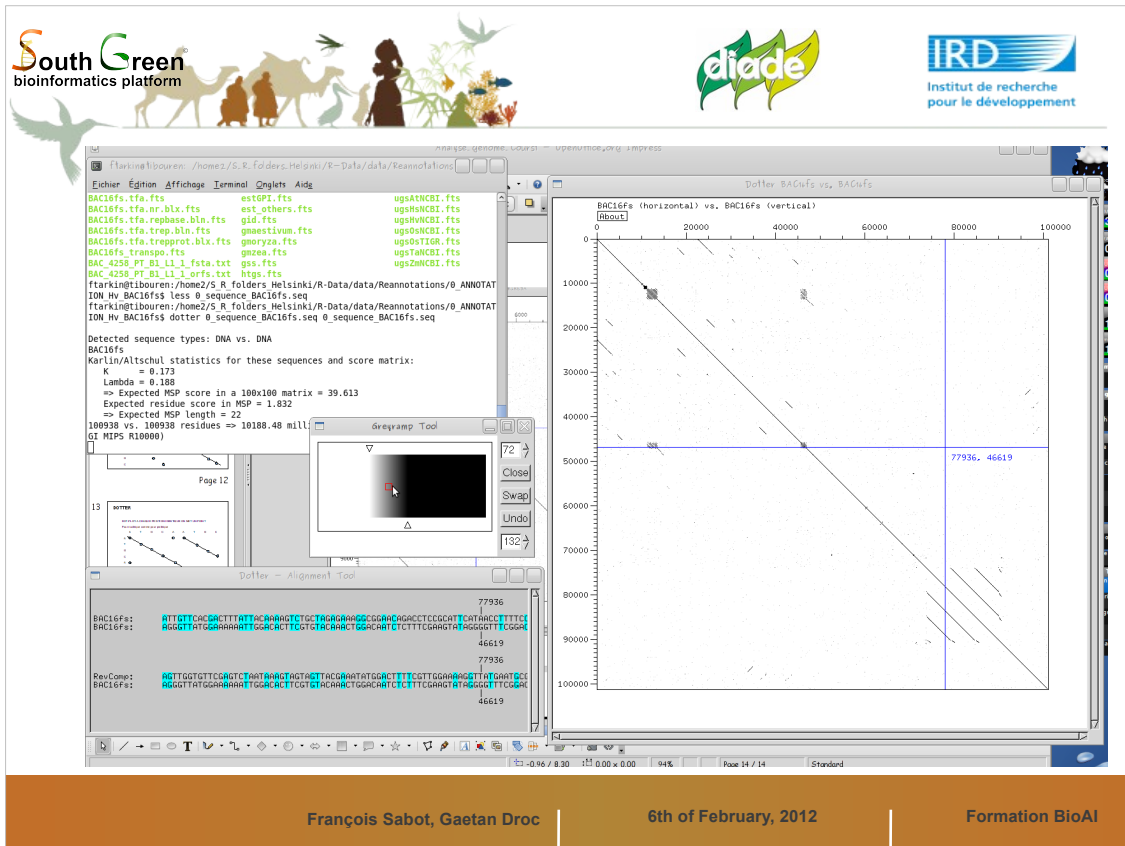


Les TIR sont très similaires, voire identiques, mais en orientation inverses. L'élément est aussi flanqué de TSD, de 2 à 19 bases

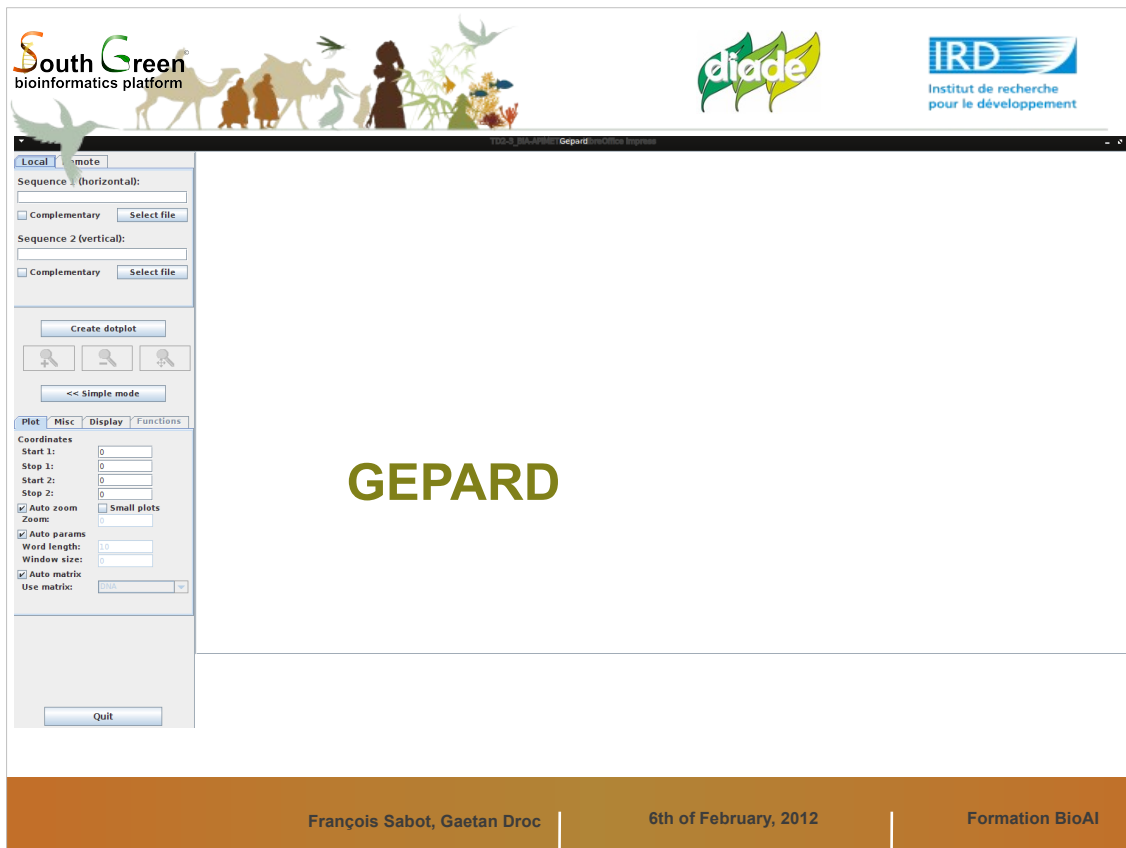


Pour déterminer les structures en répétitions directes ou indirectes, nous allons utiliser des outils de Dot-Plot.

Ici une image de Dotter, le plus souple. Mais il n'est plus disponible sur le web...

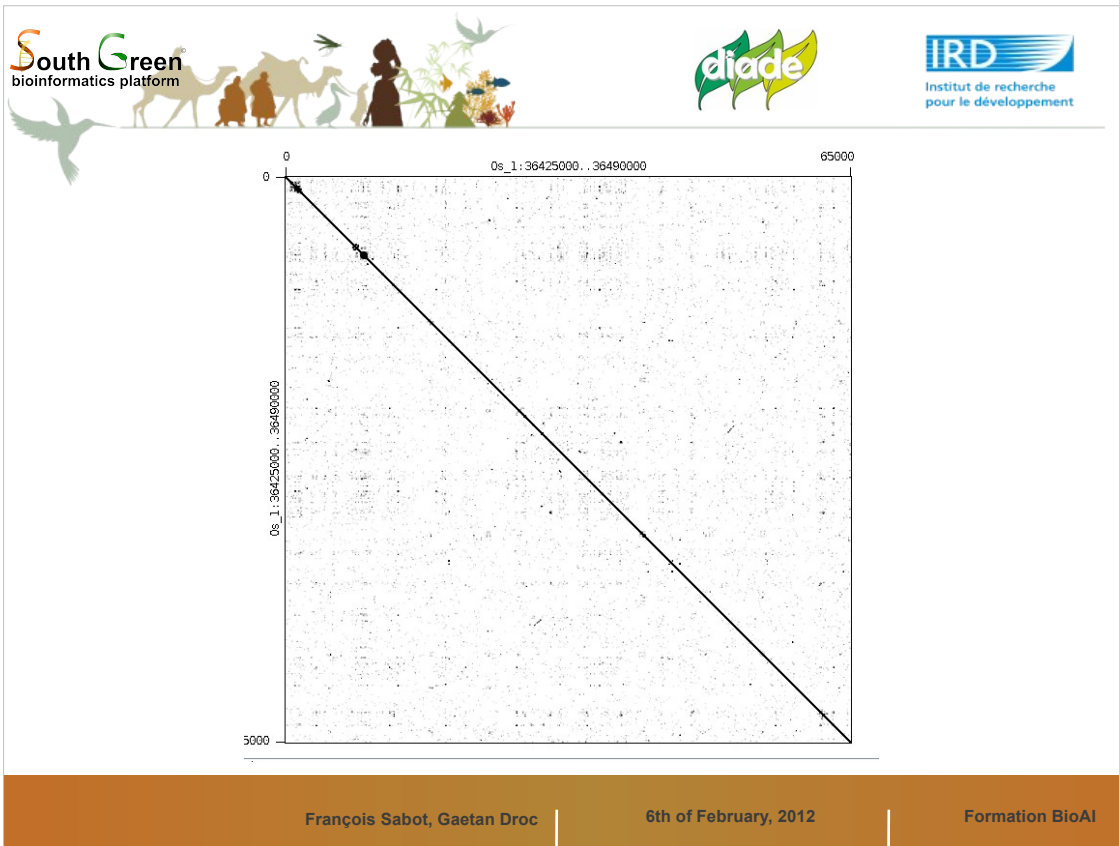


A chaque base identique le logiciel dessine un point. Donc, si une suite de bases est répétée plus loin, nous verrons deux lignes parallèles (direct) ou face-à-face (indirect, forme une croix avec la ligne centrale)

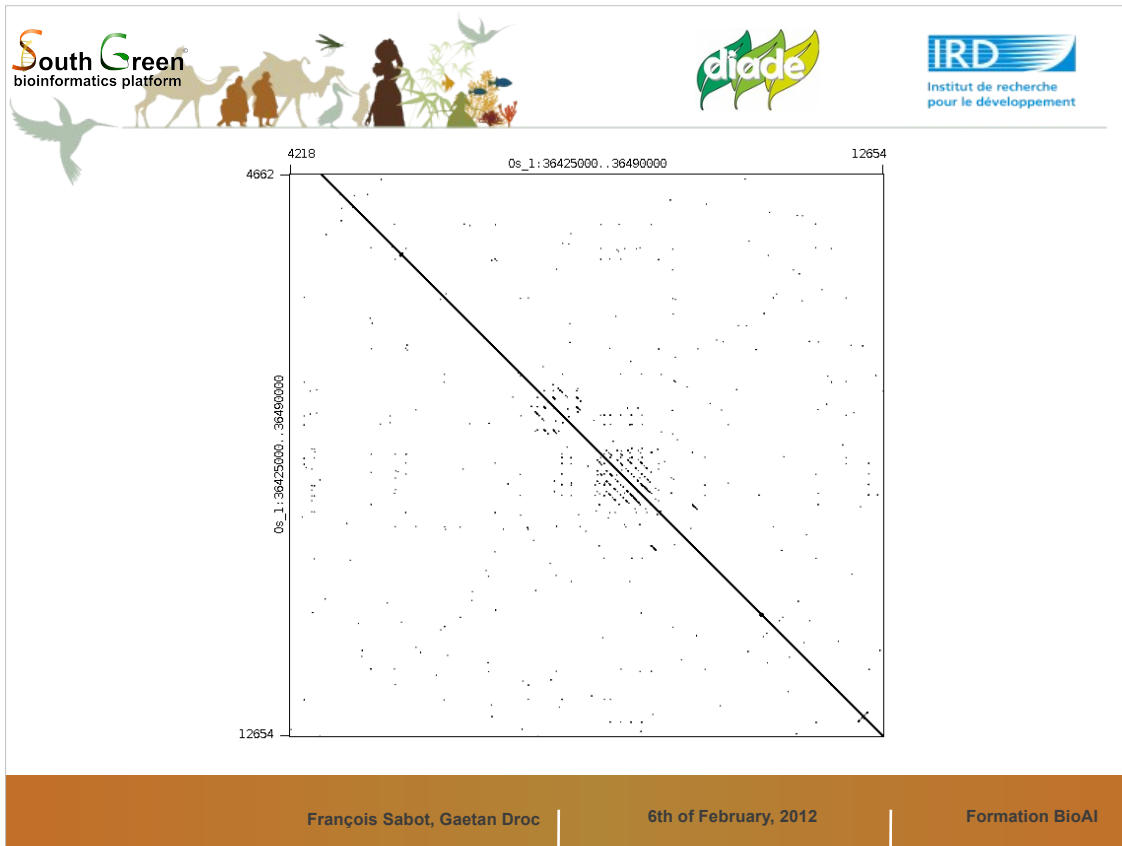


Nous allons utiliser Gepard, logiciel de dot-plot écrit en Java. Il est multiplateforme et plutôt performant.

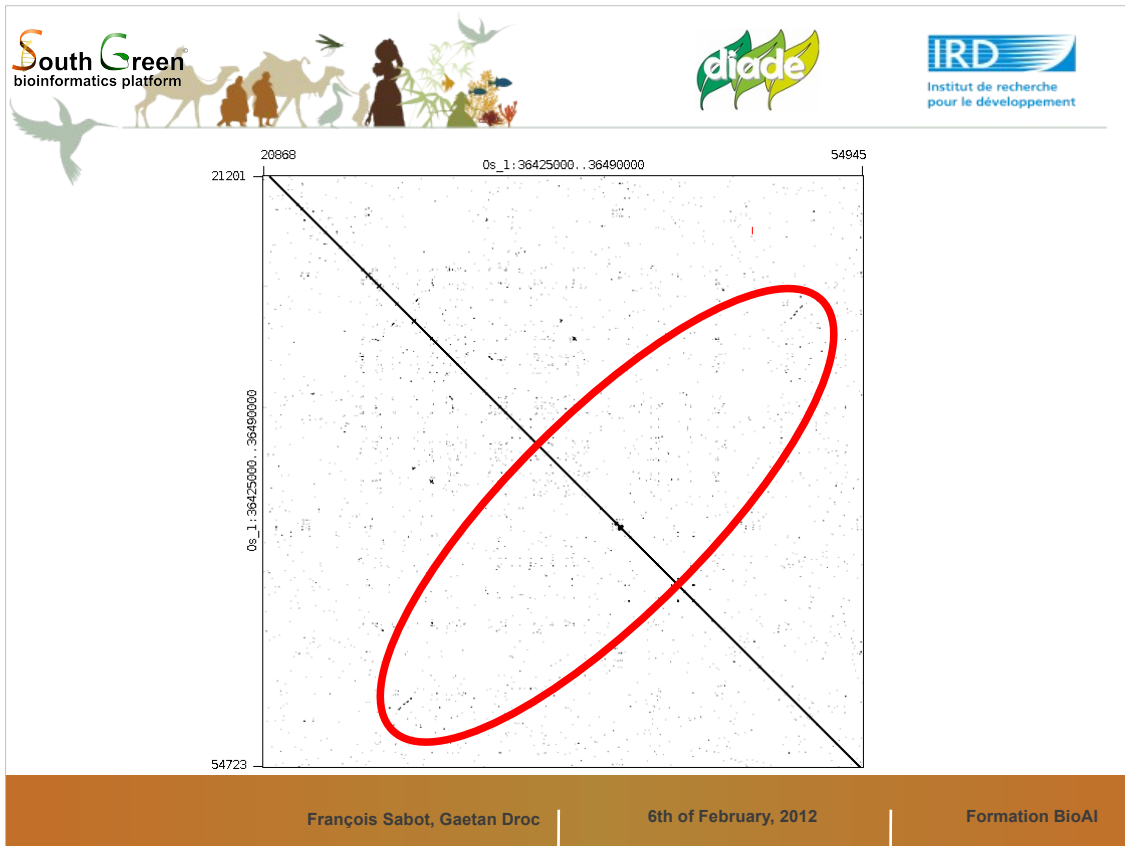
La séquence 1 et la séquence 2 sont les mêmes, il s'agit de la séquence de la Data library TE annotation récupérée ci avant. La séquence doit être de préférence en fasta.



Voici l'image de Dot d'origine. La ligne centrale représente la séquence sur elle meme.

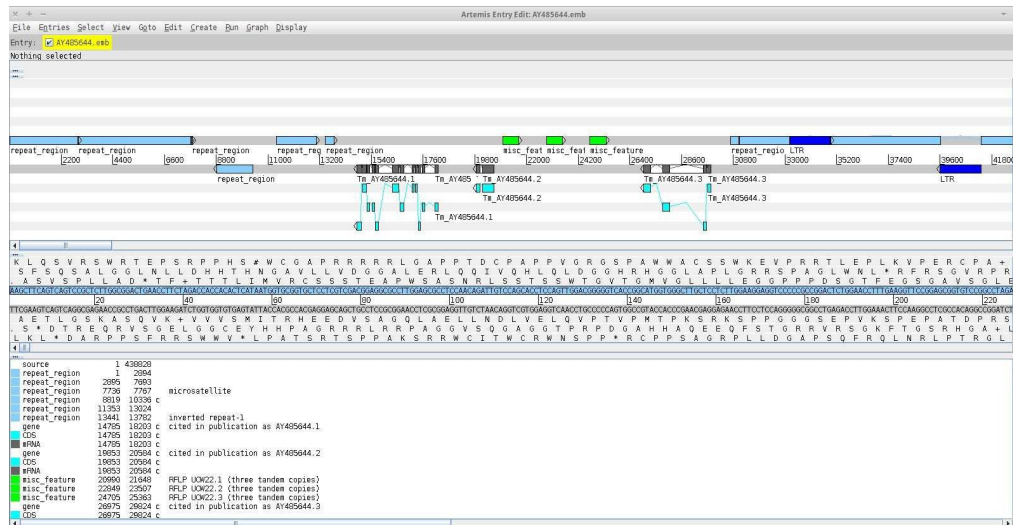


Ici une image de type microsatellite long, de type par exemple (TAGCGTA)_n.
Cela peut être aussi plusieurs répétitions courtes.

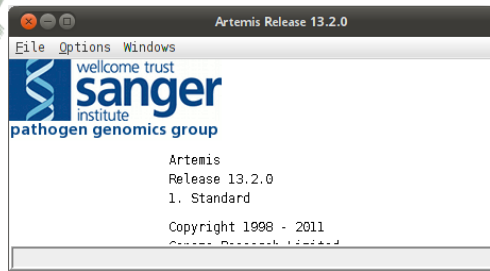


Notez ici les deux lignes en croix par rapport au centre, souvent un signe de TIR TE.

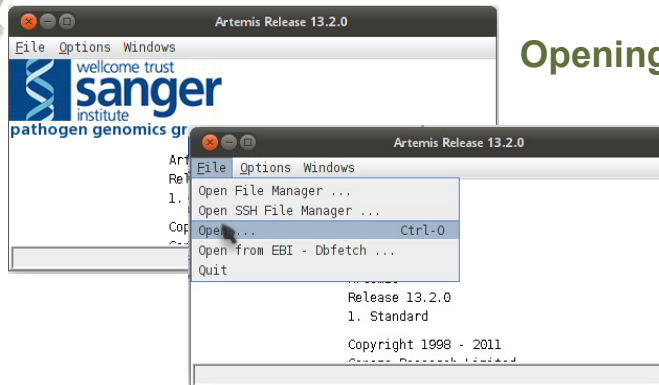
ARTEMIS: tool for viewing and annotating large sequences




<http://www.sanger.ac.uk/resources/software/artemis/>





Opening a **sequence**



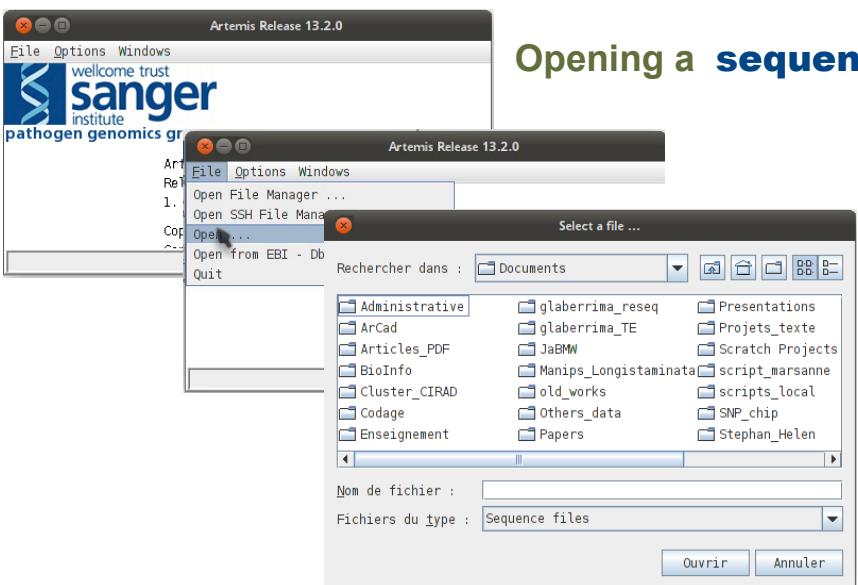
Opening a **sequence**








Opening a **sequence**





François Sabot, Gaetan Droc
6th of February, 2012
Formation BioAI

Artemis peut ouvrir des séquences Fasta, EMBL, GFF, GenBank, GTF, etc..

Si vous ne voyez pas votre fichier, demandez à voir 'Tous les fichiers' et pas seulement les 'Sequence files'.





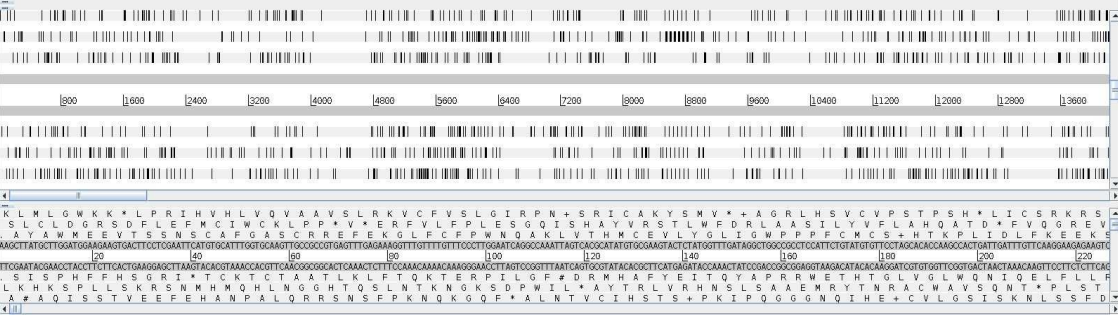


Artemis Entry Edit: annotseqDcomplete_TAA.txt

File Entries Select View Goto Edit Create Run Graph Display


Entry: ☐ annotseqDcomplete_TAA.txt

Nothing selected




François Sabot, Gaetan Droc
6th of February, 2012
Formation BioAI


Vue d'ensemble d'une séquence dans Artemis.
Les barres noires sont les codons stops



South Green
bioinformatics platform



diade




IRD
Institut de recherche
pour le développement

Adding an entry

Artemis Entry Edit: annotseq0complete_TAA.txt

File Entries Select View Goto Edit Create Run Graph Display


- Show File Manager ...
- Read An Entry ...
- Read Entry Into ...
- Read BAM ...
- Save Default Entry ...
- Save An Entry ...
- Save An Entry As ...
- Save All Entries ...
- Write ...
- Clone This Window ...
- Save As Image Files (png/jpeg) ...
- Print ...
- Print Preview ...
- Open in DNAPlotter ...
- Preferences ...
- Close




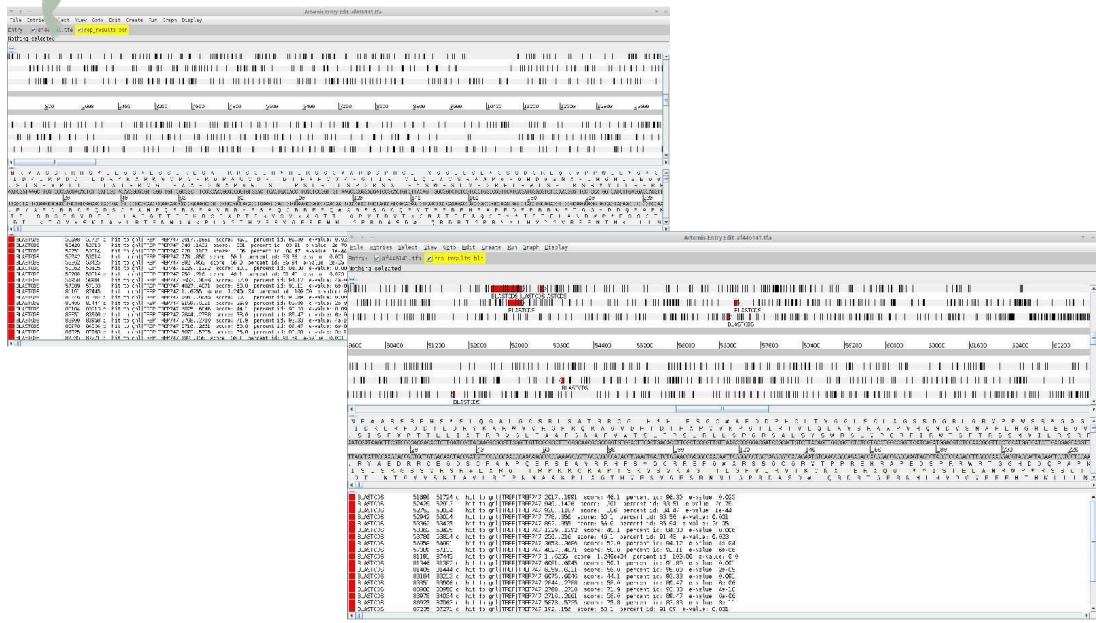
François Sabot, Gaetan Droc
6th of February, 2012
Formation BioAI

Pour ajouter une 'Entry', soit une information d'annotation/de résultat BLAST/autre, choisissez le menu comme décrit












François Sabot, Gaetan Droc

6th of February, 2012

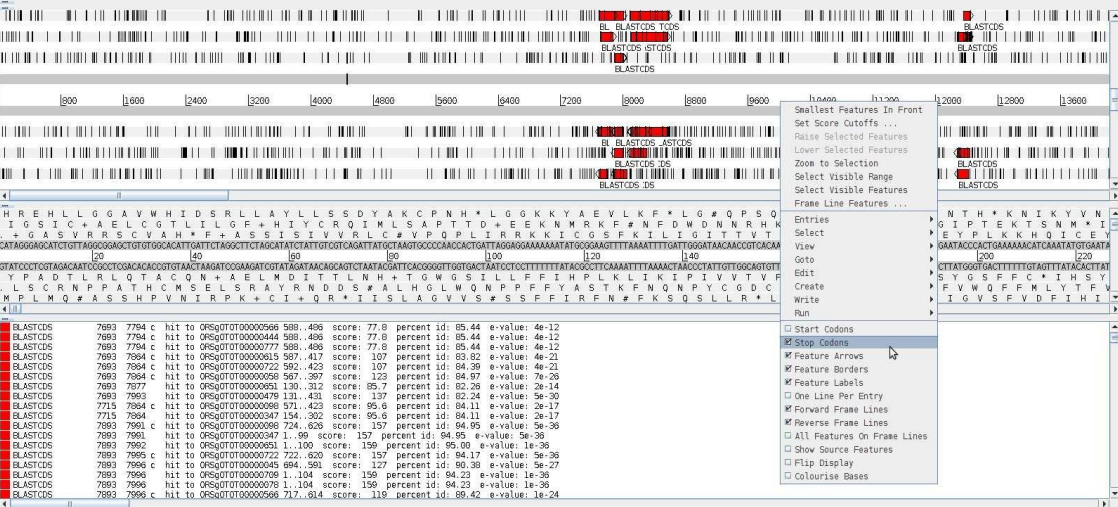
Formation BioAI

Dans le cas d'une entrée contenant un résultat BLAST (sous forme tabulée obligatoirement), les 'hits' sont notés en rouge comme BLASTCDS, meme si l'information vient d'un blast nucléaire.

Artemis Entry Edit: Os_136425000_36490000.fna

File Entry Select View Goto Edit Create Run Graph Display
Entry: ☒ Os_136425000_36490000.fna ☒ Os_136425_36490_rr-rsp-tlgr-trsp-clean.tab
3 selected bases on forward strand: 4464..4466






BLAST hit	Accession	Score	Percent ID	E-value
BLASTCD5	7693_7794 c	hit to ORSgOT0700000566 588..498	score: 77.8 percent id: 85.44 e-value: 4e-12	
BLASTCD5	7693_7794 c	hit to ORSgOT0700000444 588..498	score: 77.8 percent id: 85.44 e-value: 4e-12	
BLASTCD5	7693_7794 c	hit to ORSgOT0700000777 588..498	score: 77.8 percent id: 85.44 e-value: 4e-12	
BLASTCD5	7693_7894 c	hit to ORSgOT0700000635 567..411	score: 107 percent id: 83.82 e-value: 4e-21	
BLASTCD5	7693_7894 c	hit to ORSgOT0700000722 592..423	score: 107 percent id: 84.39 e-value: 4e-21	
BLASTCD5	7693_7894 c	hit to ORSgOT0700000656 567..397	score: 123 percent id: 84.97 e-value: 7e-26	
BLASTCD5	7693_7894 c	hit to ORSgOT0700000651 138..312	score: 85.7 percent id: 82.28 e-value: 2e-14	
BLASTCD5	7693_7893 c	hit to ORSgOT0700000476 131..431	score: 137 percent id: 82.24 e-value: 5e-30	
BLASTCD5	7715_7864 c	hit to ORSgOT0700000696 571..423	score: 85.8 percent id: 84.11 e-value: 2e-17	
BLASTCD5	7715_7864 c	hit to ORSgOT0700000347 154..302	score: 95.8 percent id: 84.11 e-value: 2e-17	
BLASTCD5	7693_7901 c	hit to ORSgOT0700000098 724..626	score: 157 percent id: 84.95 e-value: 5e-36	
BLASTCD5	7693_7901 c	hit to ORSgOT0700000347 1..89	score: 137 percent id: 84.95 e-value: 5e-36	
BLASTCD5	7693_7902 c	hit to ORSgOT0700000651 1..100	score: 159 percent id: 85.00 e-value: 1e-36	
BLASTCD5	7693_7899 c	hit to ORSgOT0700000722 722..626	score: 157 percent id: 84.17 e-value: 5e-36	
BLASTCD5	7693_7996 c	hit to ORSgOT0700000045 694..591	score: 127 percent id: 80.30 e-value: 5e-27	
BLASTCD5	7693_7998 c	hit to ORSgOT0700000708 1..104	score: 159 percent id: 84.23 e-value: 1e-36	
BLASTCD5	7693_7998 c	hit to ORSgOT0700000076 1..104	score: 159 percent id: 84.23 e-value: 1e-36	
BLASTCD5	7693_7998 c	hit to ORSgOT0700000568 717..614	score: 119 percent id: 80.42 e-value: 1e-24	

François Sabot, Gaetan Droc

6th of February, 2012

Formation BioAI

Commençons à améliorer notre visuel...



Artemis Feature Edit: misc_feature

key: misc_feature Add Qualifier: note

Location: 1..65001

Complement Grab Range Remove Range Goto Feature Select Feature TAT ObjectEdit

OK Cancel Apply

Creating a feature

François Sabot, Gaetan Droc

6th of February, 2012

Formation BioAI

En ajoutant directement via le menu Feature une nouvelle information

Artemis Feature Edit: misc_feature

Key: misc_feature Add Qualifier: note

Location: 1..65001

Complement Grab Range Remove Range Goto Feature Select Feature TAT ObjectEdit

OK Cancel

Creating a feature

Artemis Feature Edit: misc_feature

Key: misc_feature Add Qualifier: note




Location: 35..135


Complement Grab Range Remove Range Goto Feature Select Feature TAT ObjectEdit

/label=test

OK Cancel Apply

Formation BioAI



BL BLASTCDS _ASTCDS

BLASTCDS

Creating from base range

```

K K Y A E V L K F * L G # Q P S Q K I N K S L E L I M G N T H * K N I K
K N M R K F # N F D W D N N R H K R # T S L L N * # W G I P T E K T S N
K I C G S F K I L I G I T T V T K D K Q V S * I D N G E Y P L K K H Q :
A A A A T A T G S G G A A G T T T A A A A T T T G A T T G G G A T A A C A A C C G T C A C A A A A G A T A A C A A G T C T C T T G A A T T G A T A A T G G G G A A T A C C C A C T G A A A A A C A T C A A A T
120 140 160 180 200
T T T T T A T A C G C C T T C A A A A T T T T A A A A C T A A C C C T A T T G T T G G C A G T G T T T C T A T T T G T T C A G A G A A C T T A A C T A T T A C C C C T T A T G G G T G A C T T T T T G T A G T T T
F I H P L K L I K I P I V V T V F S L C T E Q I S L P S Y G S F F C * I
F Y A S T K F N Q N P Y C G D C F I F L D R S N I I P F V W Q F F M L
F F I R F N # F K S Q S L L R * L L Y V L R K F Q Y H P I G V S F V D F
19
8-12
12 _

```

François Sabot, Gaetan Droc

6th of February, 2012

Formation BioAI

En sélectionnant une zone à la main



BL BLASTCDS _ASTCDS

BLASTCDS

Creating from base range

```

K K Y A E V L K F * L G [ # Q P S O K I N K S L E L I M G N ] T H * K N I K
K N M R K F # N F D W D N N R H K R # T S L L N * # W G I P T E K T S N
K I C G S F K I L I G I T T V T K D K Q V S * I D N G E Y P L K K H Q :
A A A A T A T G S G A A G T T T A A A A T T G A T T G G G A T A A C A A C C G T C A C A A A G A T A A C A A G T C T C T T G A A T T G A T A A T G G G A A T A C C C A C T G A A A A A C A T C A A A T
120 140 160 180 200
T T T T T A C G C C T T C A A A A T T T A A A A C T A A C C C T A T T G T T G C C A G T G T T T C T A T T T G T T C A G A G A C T T A A C T A T T A C C C C T T A T G G G T G A C T T T T T G T A G T T T
F I H P L K L I K I P I V V T V F S L C T E Q I S L P S Y G S F F C * I
F Y A S T K F N Q N P Y C G D C F I F L D R S N I I P F V W Q F F M L
F F I R F N # F K S Q S L L R * L L Y V L R K F Q Y H P I G V S F V D F

```

Artemis Entry Edit: Os_1:36425000_36490000.fna

File Entries Select View Goto Edit Create Run Graph Display

Entry: Os_1-36425000_36490000.fna

51 selected bases on forward strand: 145

Feature From Base Range Ctrl-C

Intron Features

Intergenic Features

Exon Features

Gene Features

New Entry

Mark Open Reading Frames ...

Mark Empty ORFs ...

Mark ORFs In Range ...

Mark From Pattern ...

Mark Ambiguities



BL BLASTCDS TCDS

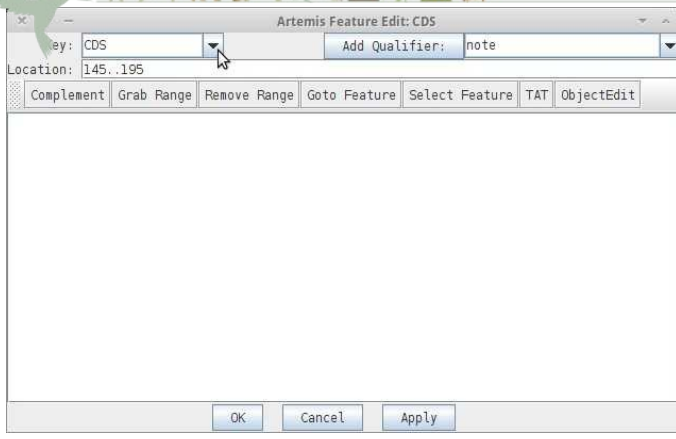
BLASTCDS 1

BL BLASTCDS _ASTCDS

BLASTCDS

H R E H L L G G A V W H I D S R L L A Y L L S S D Y A K C P N H * L G G K K Y A E V L K F * L G [# Q P S O K I N K S L E L I M G N] T H * K N I K Y V N
I G S I C + A E L C G T L I L G F + H I Y C R Q I M L S A P T T D + E E K N M R K F # N F D W D N N R H K R # T S L L N * # W G I P T E K T S N M * I
+ G A S V R R S C V A H * P + A S S I S I V V R L C # V P Q P L I R K K I C G S F K I L I G I T T V T K D K Q V S * I D N G E Y P L K K H Q I C E V
C A T A G S G A C A T C T G T T A G G G S A C T G T G T G G A T T G A T T G G G A T A A C A A C C G T C A C A A A G A T A A C A A G T C T C T T G A A T T G A T A A T G G G A A T A C C C A C T G A A A A A C A T C A A A T G A T A
20 40 60 80 100 120 140 160 180 200 220
G A T A C C C T G T A G A C A T C G G C T G S A C A C C G T G A C T A G A T C G A G A T C A T A G A T T C A C G S G G T G T G A C T A A T C C G T T T T T A T A G G C T C A A A T T T A A A C T A A C C C T A T T G T T G C C A G T G T T T C T A T T T G T T C A G A G A C T T A A C T A T T A C C C C T T A T G G G T G A C T T T T T G T A G T T T A C A C T T A
Y P A D T L R L Q T A C A N + A E L M O I T T L N H + T G W G S T L L F F I H P L K L I K I P I V V T V F S L C T E Q I S L P S Y G S F F C * I H S Y
L S C R N P P A T H C M S E L S R A Y R N D D S * A L H G L W O N P P F F Y A S T K F N Q N P Y C G D C F I F L D R S N I I P F V W Q F F M L Y T F V
W P L M Q # A S S H P V N I R P K + C I + Q R * I I S L A G V S # S S F F I R F N # F K S Q S L L R * L L Y V L R K F Q Y H P I G V S F V D F I H I
* I I I





Artemis Feature Edit: CDS

Key: CDS Add Qualifier: note

Location: 145..195

Complement Grab Range Remove Range Goto Feature Select Feature TAT ObjectEdit

OK Cancel Apply

Changing key & qualifier

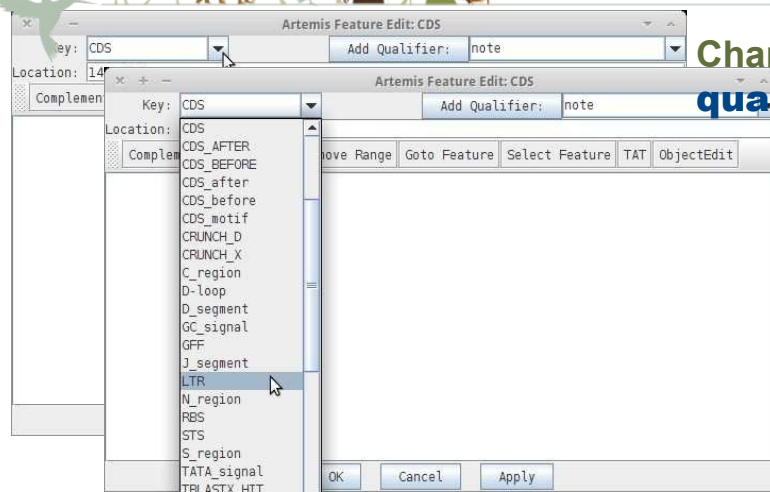
François Sabot, Gaetan Droc

6th of February, 2012

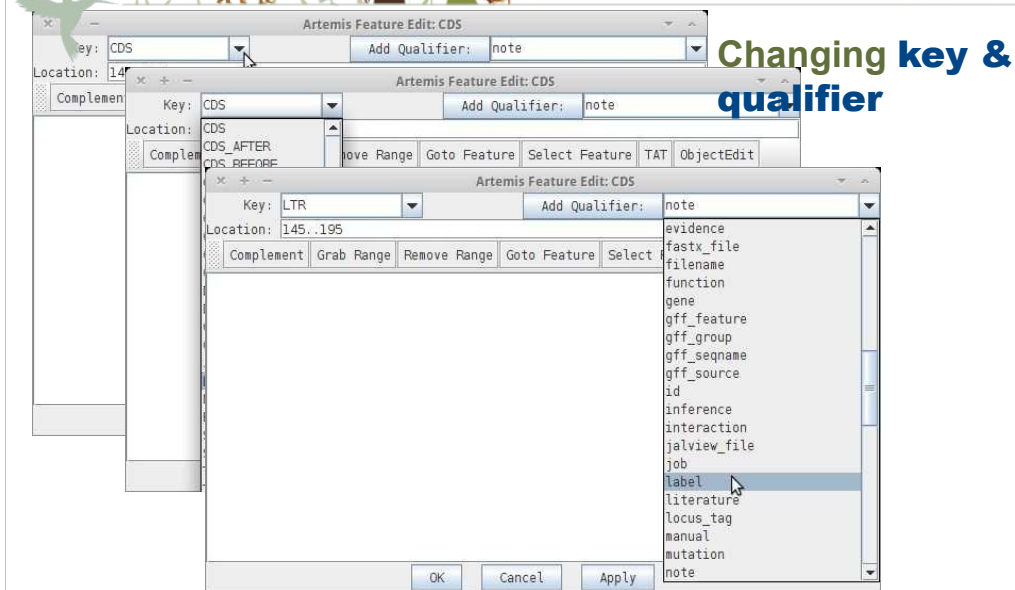
Formation BioAI

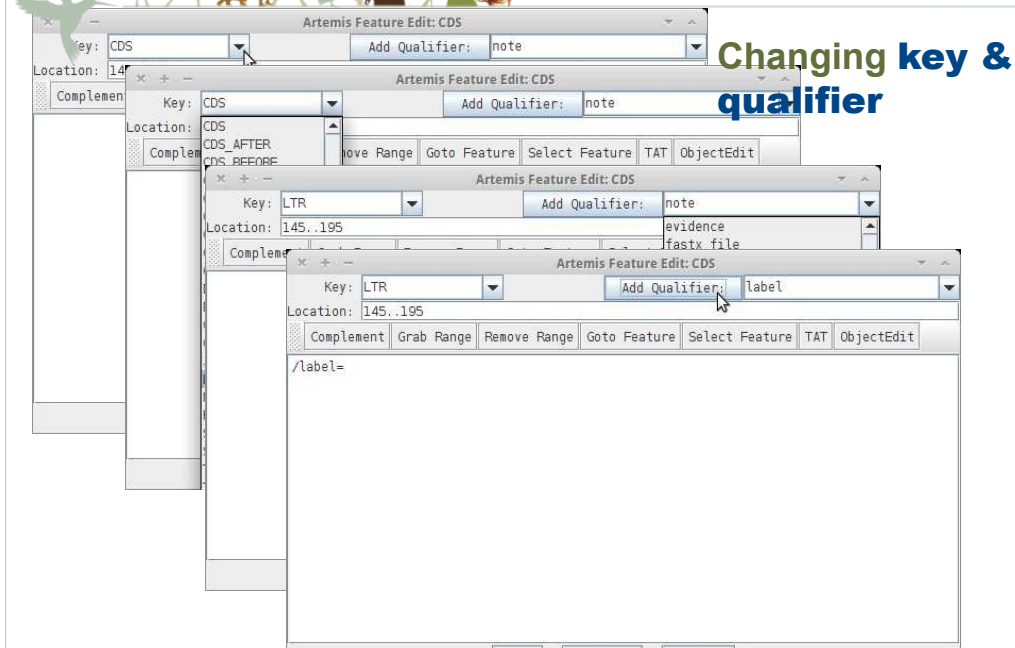
On peut aussi modifier le type de feature et lui apporter des qualificatifs.

On peut modifier une feature en la sélectionnant, et en tapant Ctrl + E, ou bien via le menu Feature/Edit selected feature in Editor




Changing key &
qualifier






Changing key & qualifier


Changing key & qualifier



South Green
bioinformatics platform



diade




IRD
Institut de recherche
pour le développement

Artemis Entry Edit: Sequence.fasta

File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Sequence.fasta ☒ repeats.bln ☒ no name

Selected feature: bases 704 BLASTCDS (/blast score= 345 /score=81.23 /percent id=81.23 /query id=0s 1:36426001...36558000 /subject start=11950 /sub)



Copying feature

Smallest Features In Front

Set Score Cutoffs ...

Raise Selected Features

Lower Selected Features

Zoom to Selection

Select Visible Range

Select Visible Features

Frame Line Features ...

Entries

Select

View

Goto

Undo

Ctrl-U

Redo

Ctrl-E

Find/Replace Qualifier Text ...

Qualifier of Selected Feature(s)

Selected Feature(s)

Move Selected Features To

Copy Selected Features To

Trim Selected Features

Extend Selected Features

Automatically Create Gene Names

Fix Gene Names

Fix Stop Codons

Bases

Contig Reordering

Header Of Default Entry

François Sabot, Gaetan Droc
6th of February, 2012
Formation BioAI

Puis sélectionnez Entry/Copy entry to
Choisissez le nom de votre entrée.

File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Sequence.fasta ☒ repeats.blm ☒ no name

Copying feature

Selected feature: bases 704 BLASTCD5 (/blast score= 345 /score=81.23 /percent id=81.23 /query id=0s 1:36429001..36558000 /subject start=11950 /sub



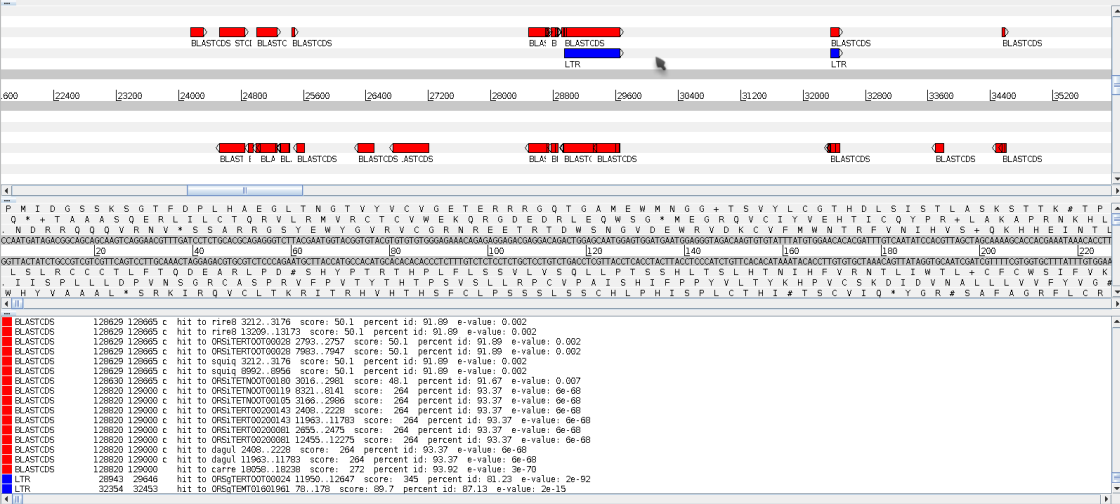
Artemis Entry Edit: Sequence.fasta


File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Sequence.fasta ☒ repeats.blm ☒ no name


Nothing selected

Merging features






South Green
bioinformatics platform



diade



IRD
Institut de recherche
pour le développement

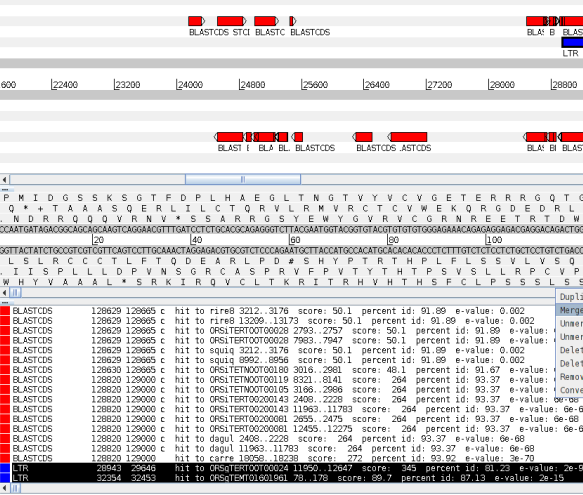
Merging features

Artemis Entry Edit: Sequence.fasta

File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Sequence.fasta ☒ repeats.bln ☒ no name

2 selected features _total bases 804 (LTR.LTR)



Smallest Features In Front

Set Score Cutoffs ...

Raise Selected Features

Lower Selected Features

Zoom to Selection

Select Visible Range

Select Visible Features

Frame Line Features ...

Entries

Select

View

Goto

Edit

Create

Write

Start Codons

Stop Codons

Duplicate

Merge

Unmerge

Unmerge All Segments

Delete

Delete Exons

Remove Introns

Convert Keys ...

Flip Display

Colourise Bases

Undo

Redo

Selected Features in Editor

Subsequence (and Features)

Find/Replace Qualifier Text ...

Qualifier of Selected Feature(s)

Selected Feature(s)

Move Selected Features To

Copy Selected Features To

Trim Selected Features

Extend Selected Features

Fix Stop Codons

Automatically Create Gene Names

Fix Gene Names

Bases

Contig Reordering

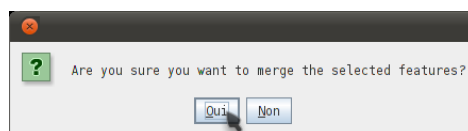
Header Of Default Entry

François Sabot, Gaetan Droc
6th of February, 2012
Formation BioAI

On peut aussi joindre ('Merge') deux features.
Sélectionnez les deux (ou plus), puis
clic-droit/Entry/Merge.

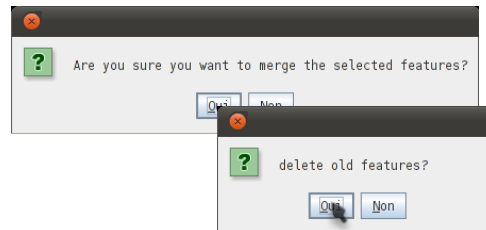
Ou bien sélectionnez les, puis Ctrl + M

Merging features





Merging features



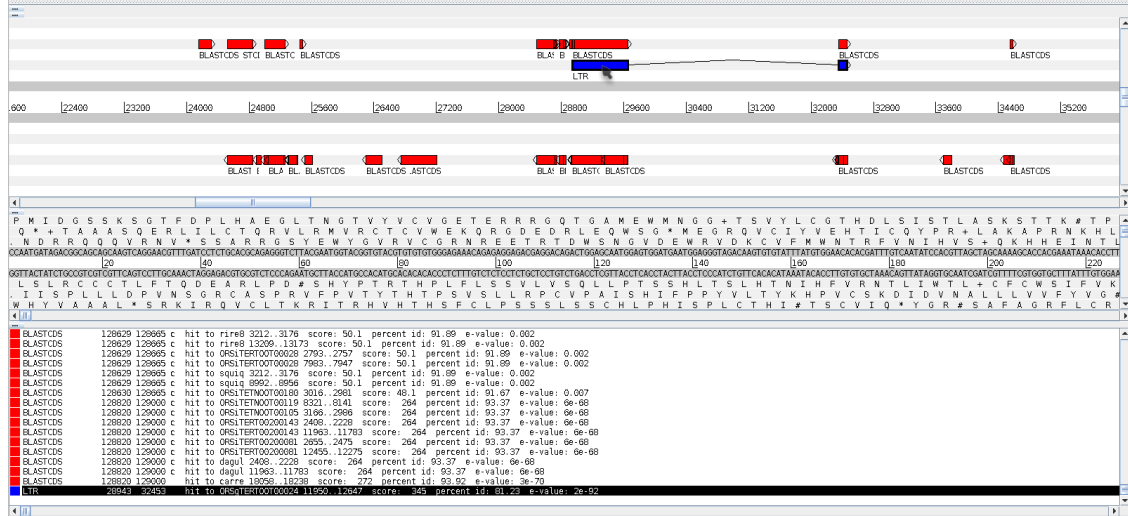
Artemis Entry Edit: Sequence.fasta


File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Sequence.fasta ☒ repeats.blm ☒ no name


Selected feature: bases 804 LTR (/blast score= 345/blast score=89.7 /score=81.23/score=87.13 /percent id=81.23/percent id=87.13 /query id=0s 1:3642

Merging features






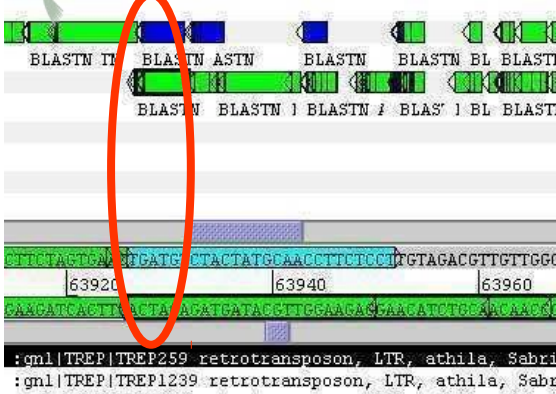
South Green
bioinformatics platform

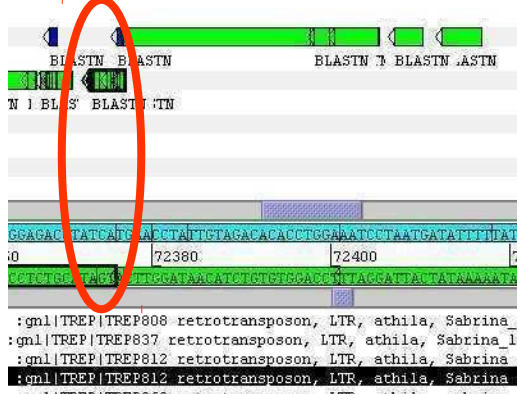


diade



IRD
Institut de recherche
pour le développement





This element starts in 72375 and ends in 63926
(negative strand), first and last “logical” hits for
***Sabrina*, Gypsy, LTR retrotransposon**

François Sabot, Gaetan Droc
6th of February, 2012
Formation BioAI

Quelques informations supplémentaires...



Manual identification of the TSD

Sabrina_AF446141-1

63940 63960

ACTTG

-Sabrina_AF446141-1 /rpt type=Class I, LTR retrotransposon

Sabrina_AF446141-1

72340 72360

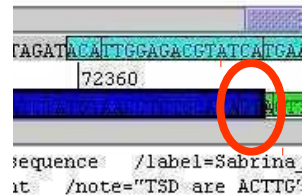
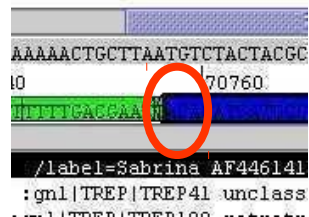
ACTTG

72375 c Complete element /label=Sabrina_AF446141-1

TSD are **ACTTG** here, and are added to the description



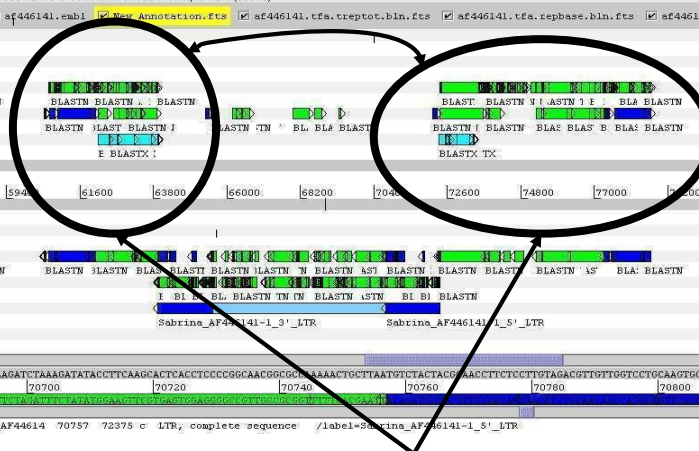
Identification of the 5' LTR TG..CA



5' LTR from 70756..72375, new feature created

Key feature: LTR

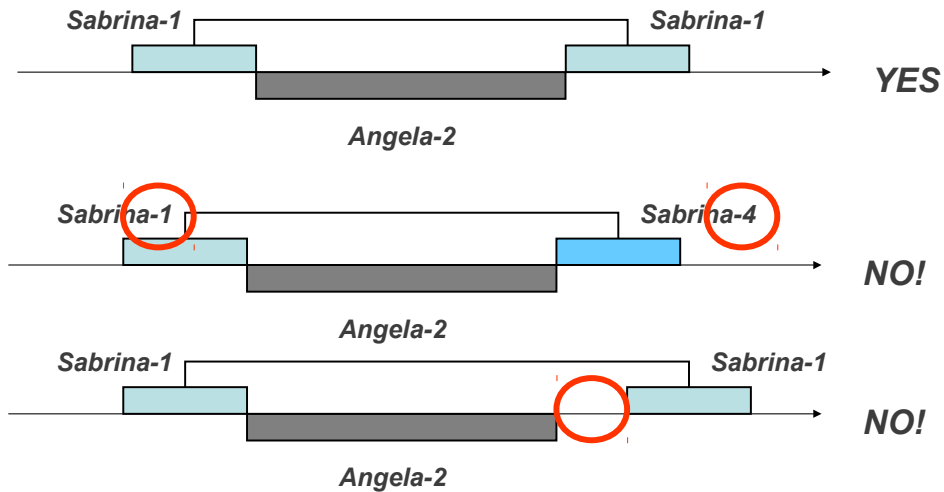
Complete element



Classical aspect of a **nested element**: same hits on each side,
Continuous and related (no breaks in the sequence)

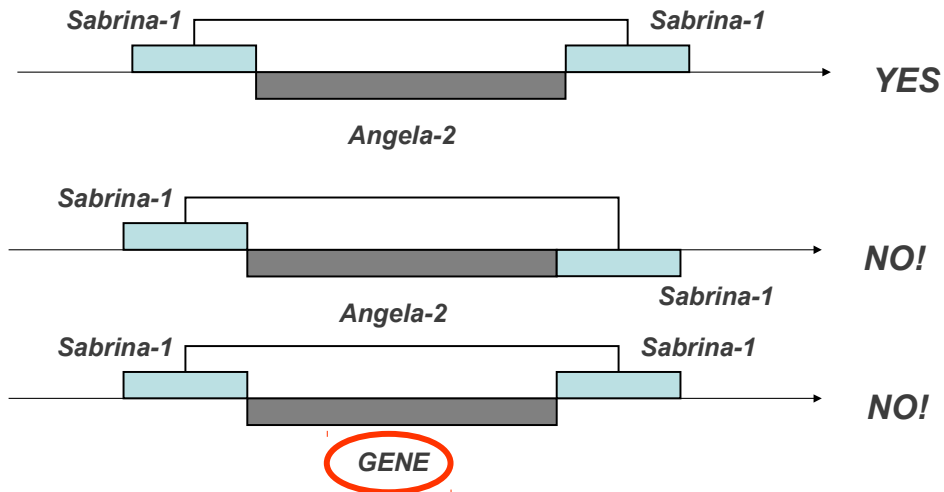


Be careful for the **merging** of elements! Some examples





Be careful for the **merging** of elements! Some examples





Few additional **advices**

- _ Always prefer the **BLASTn** to the BLASTx results
- _ Always prefer the **Specific** db to the RepBase (more details)
- _ Be careful **not to cross multiple** annotations at the same position (e.g., end of element X and start of element y at the same nucleotide)
- _ Do not fear to put “**uncertain**” as a comment for an ambiguous annotation
- _ An EST/cDNA is **not necessary** a gene