Introduction à la bioinformatique appliquée à la génomique

Christine Tranchant-Dubreuil

IRD

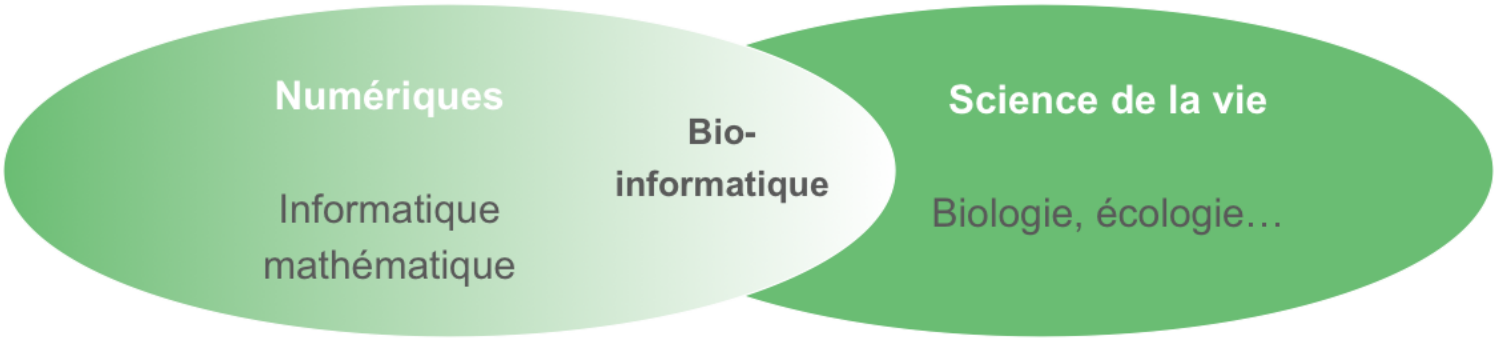
Ecole Rhématique Régionale en bioinformatique
Dakar, 4-8 Novembre 2013

Pourquoi la bioinformatique ?

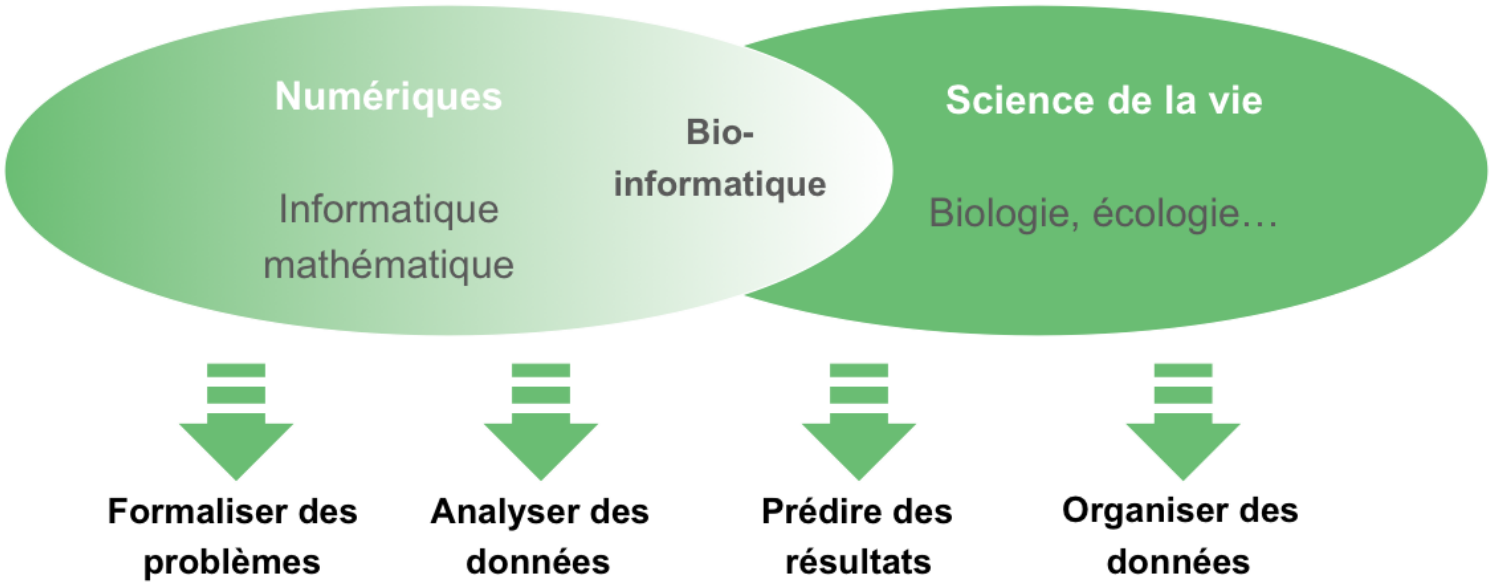
- Un nombre croissant de projets de séquençage
- Une masse de données génomiques publiques en croissance constante : banques de données nucléiques (ESTs, BACs), protéiques, génome etc.
- De nombreuses ressources web disponibles :
  - Définition *in silico* de marqueurs de type SSR, SNP, élément transposable
  - Rechercher un gène d'intèrêt séquencé dans des espèces proches
  - Annotation d'un BAC contenant un gène d'intérêt
  - Génomique comparative

Une expertise en bioinformatique indispensable pour analyser et exploiter cette véritable masse d'informations génomiques librement accessible.

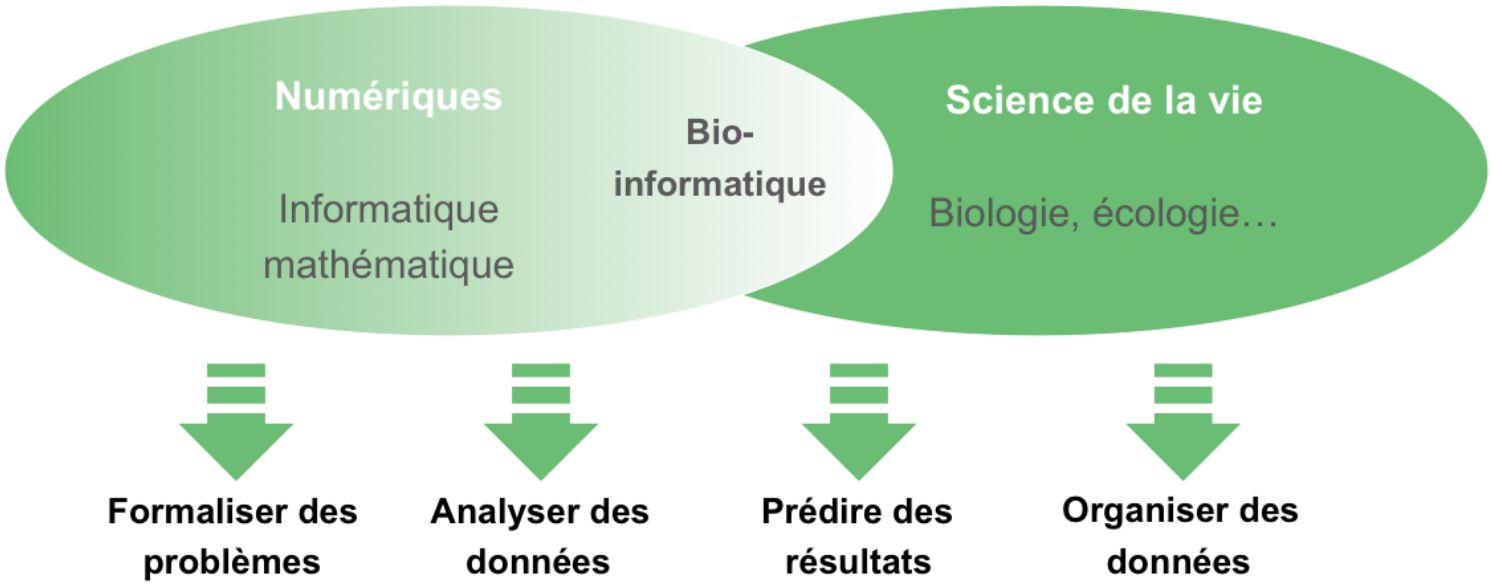
Qu'est ce que la bioinformatique ?



Qu'est ce que la bioinformatique ?



Qu'est ce que la bioinformatique ?



Algorithmique, génie logiciel, plateforme de calcul

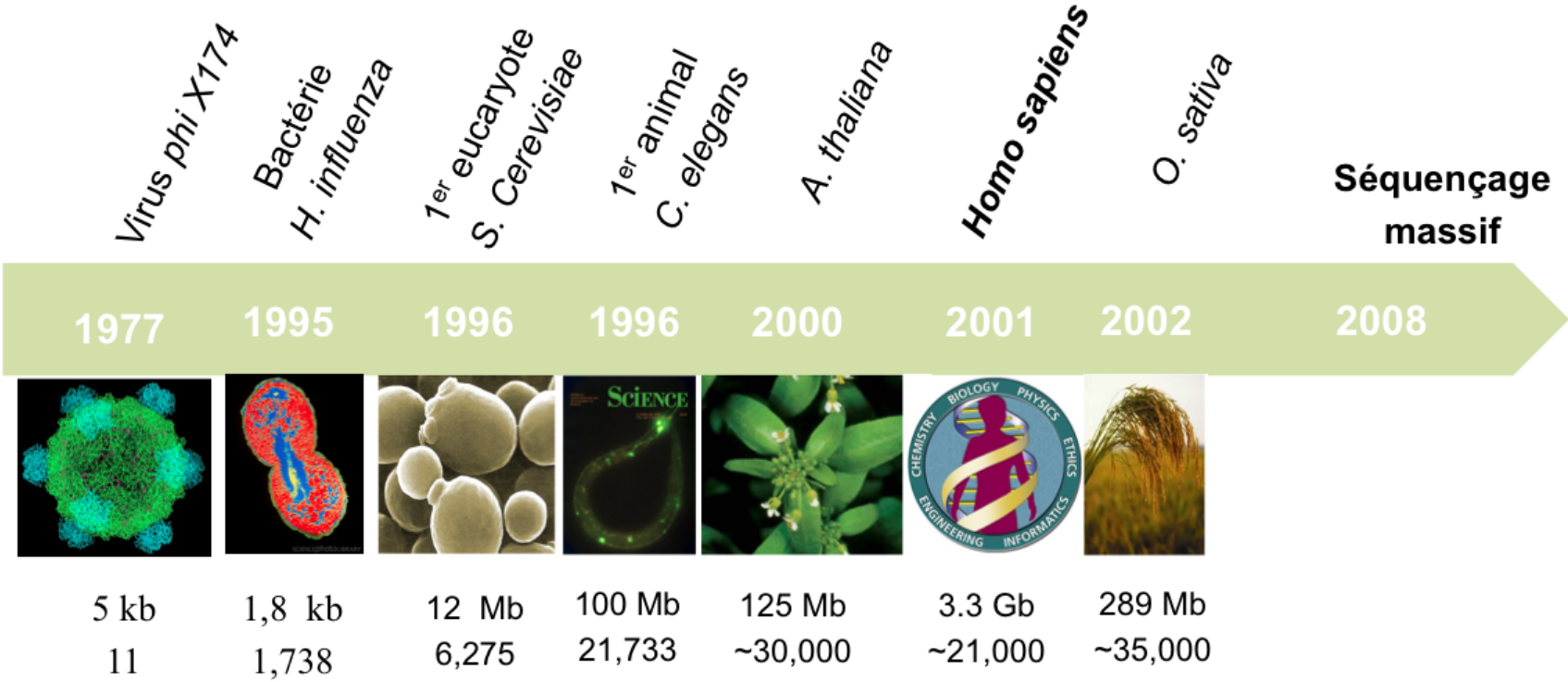
Génomique et bioinformatique ?

In silico vs *in papyro* ?

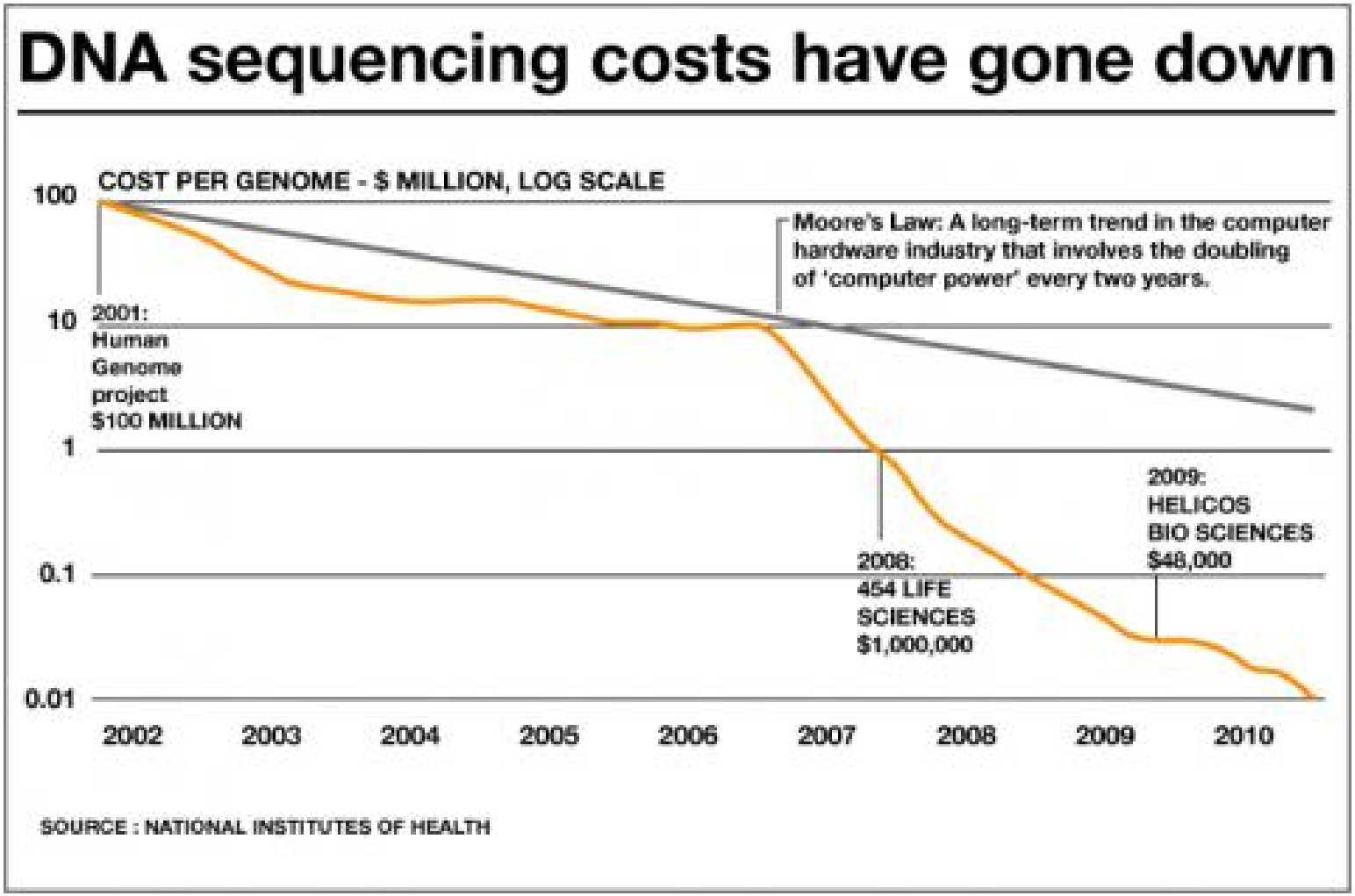
- **Génomique** : Etude des génomes et de l'ensemble de leurs gènes (structure, expression et régulation des gènes, évolution et dynamique)
- **Bioinformatique** : Approche *in silico* de la biologie indispensable pour :
  - Stocker et gérer les données de séquençage des génomes
  - Leur donner un sens

Incontournable pour exploiter les volumes de données issus des projets de séquençage.

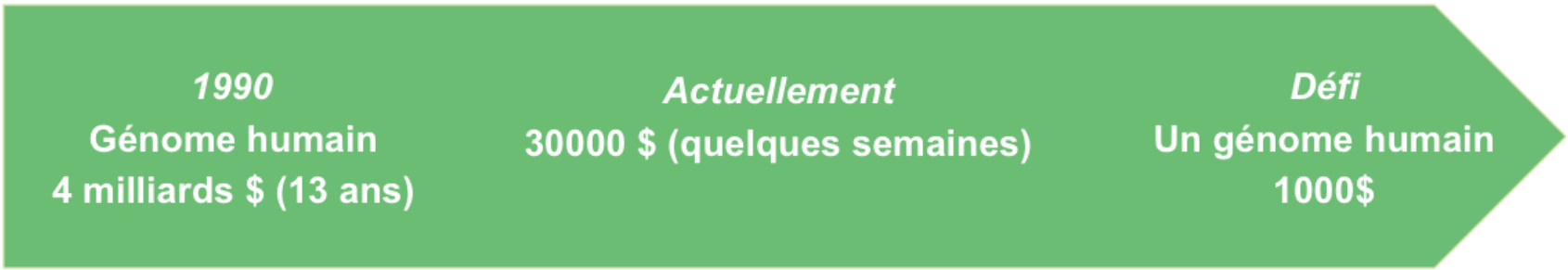
Evolution du séquençage des génomes



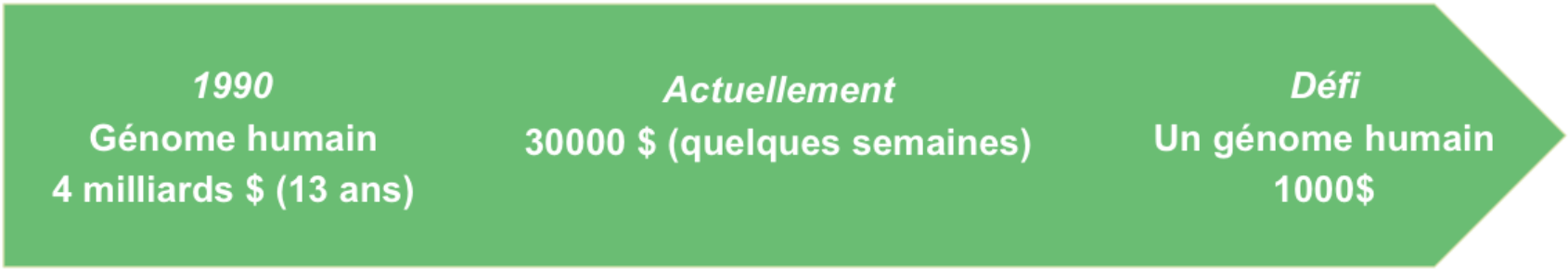
Evolution du séquençage des génomes



Révolution des technologies de séquençage

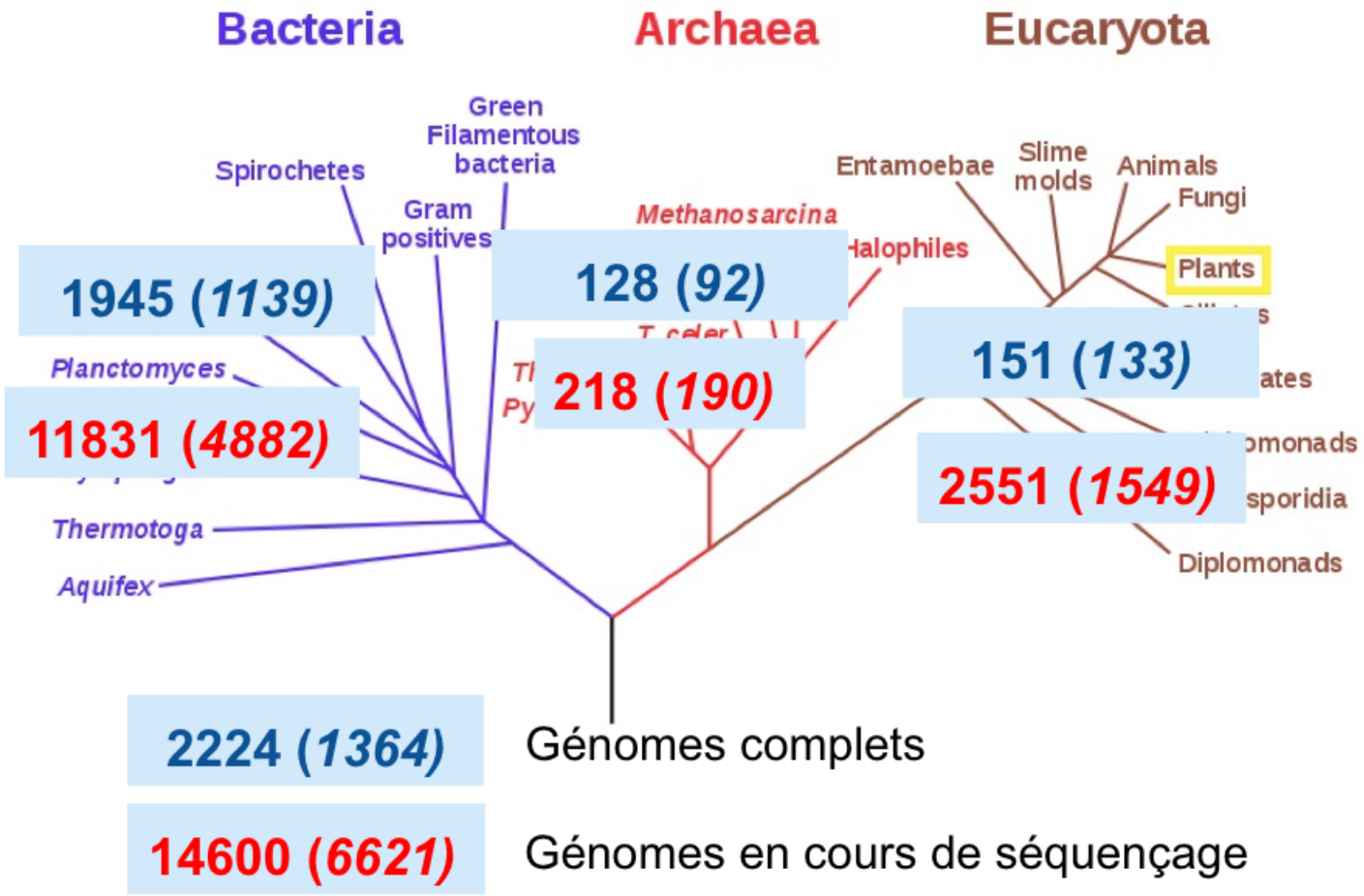


Révolution des technologies de séquençage



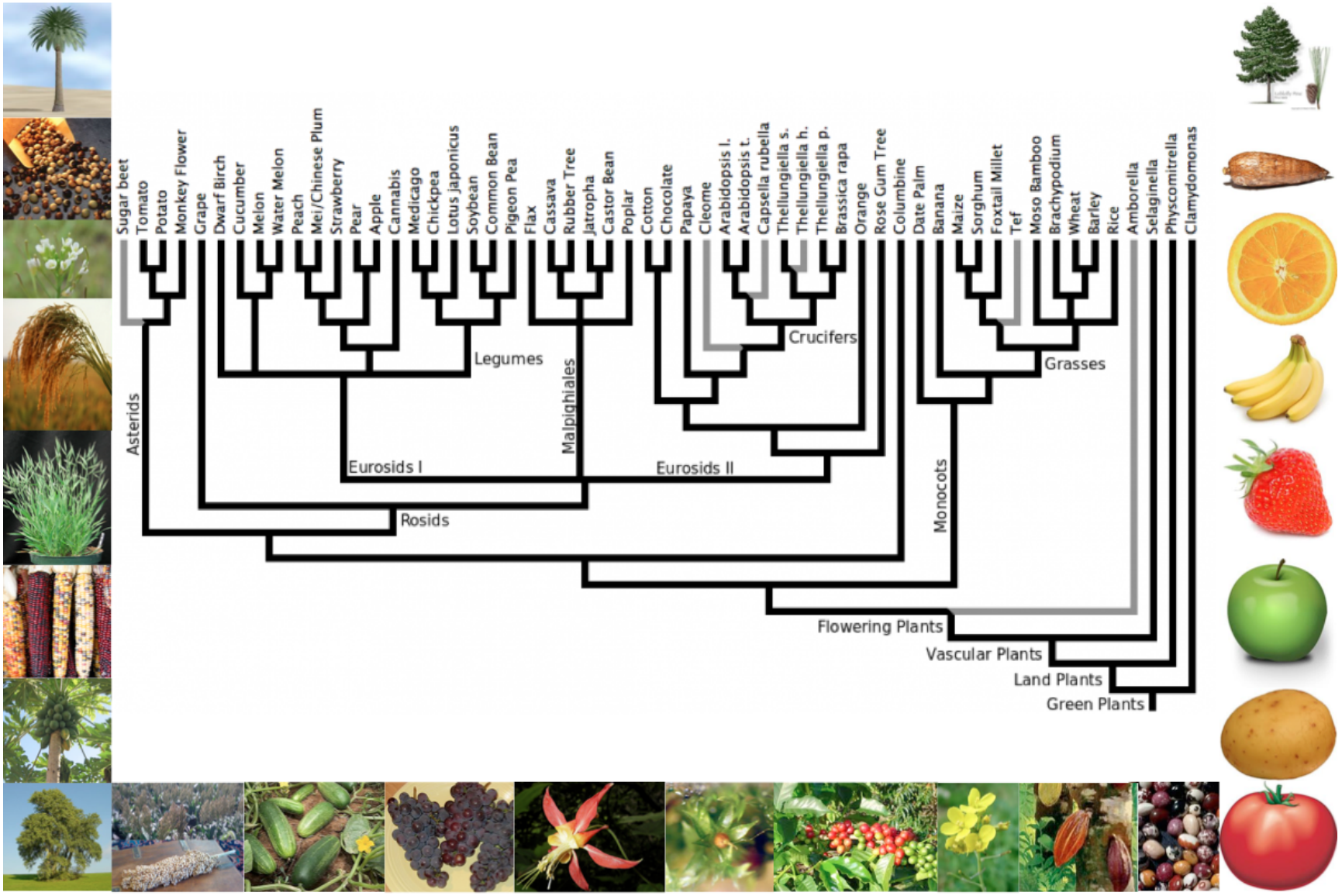
Une explosion des projets de séquençage et des données génomiques

Bilan des projets génomes en 2012 (2010)

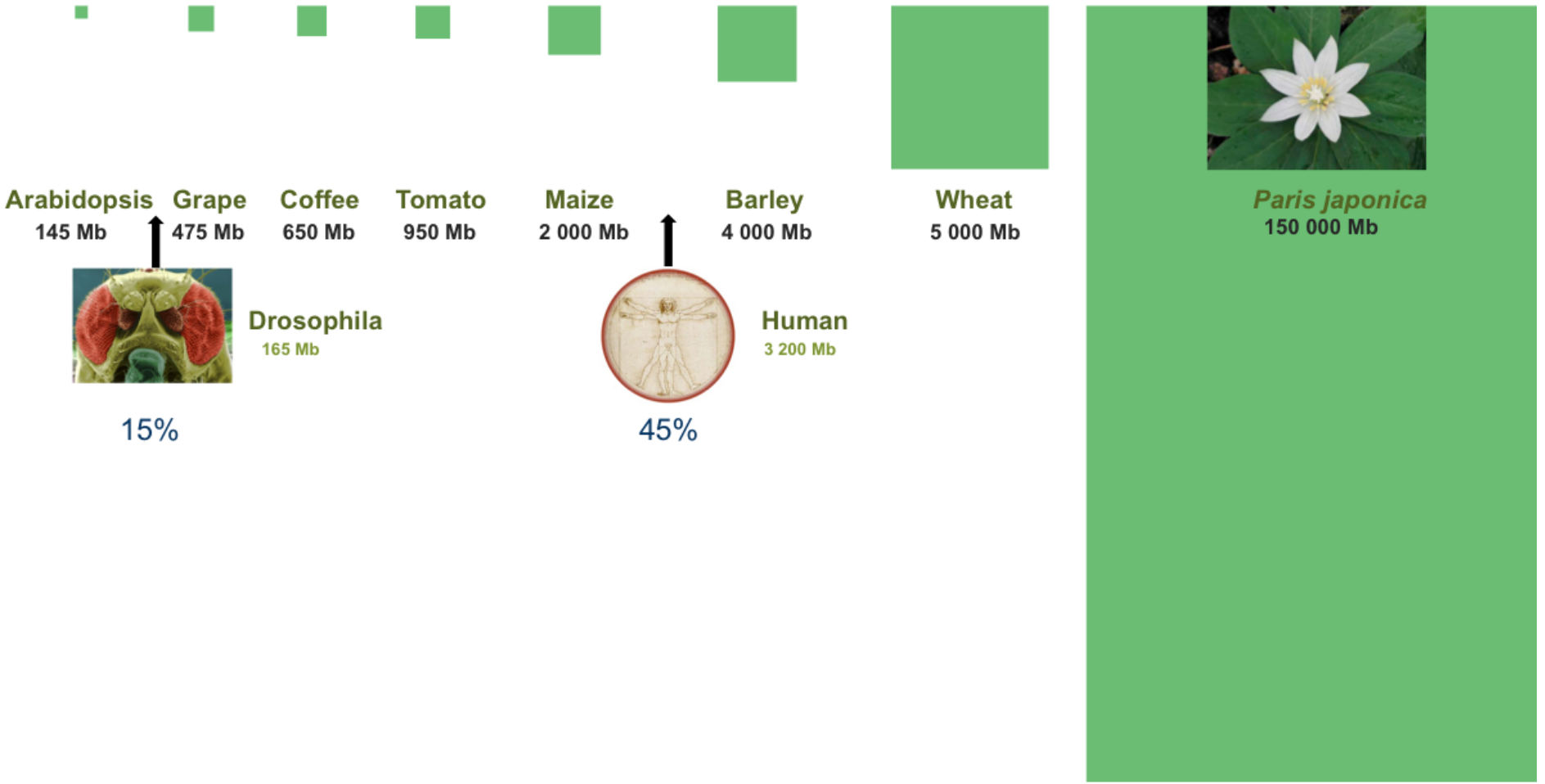


http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html
http://www.genomesonline.org

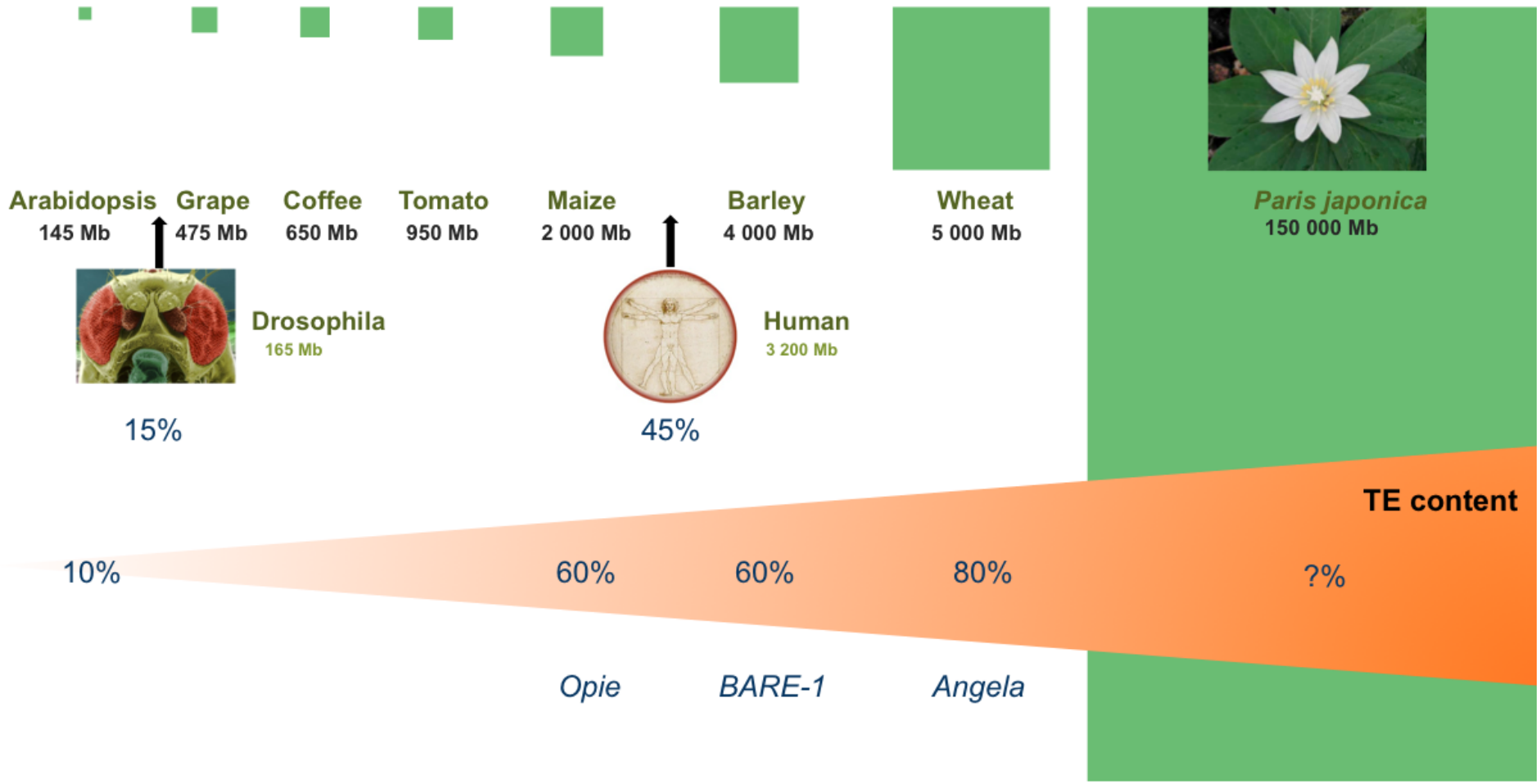
Arbre phylogénétique des génomes de plantes séquencés



Le paradoxe de la C-value



Le paradoxe de la C-value



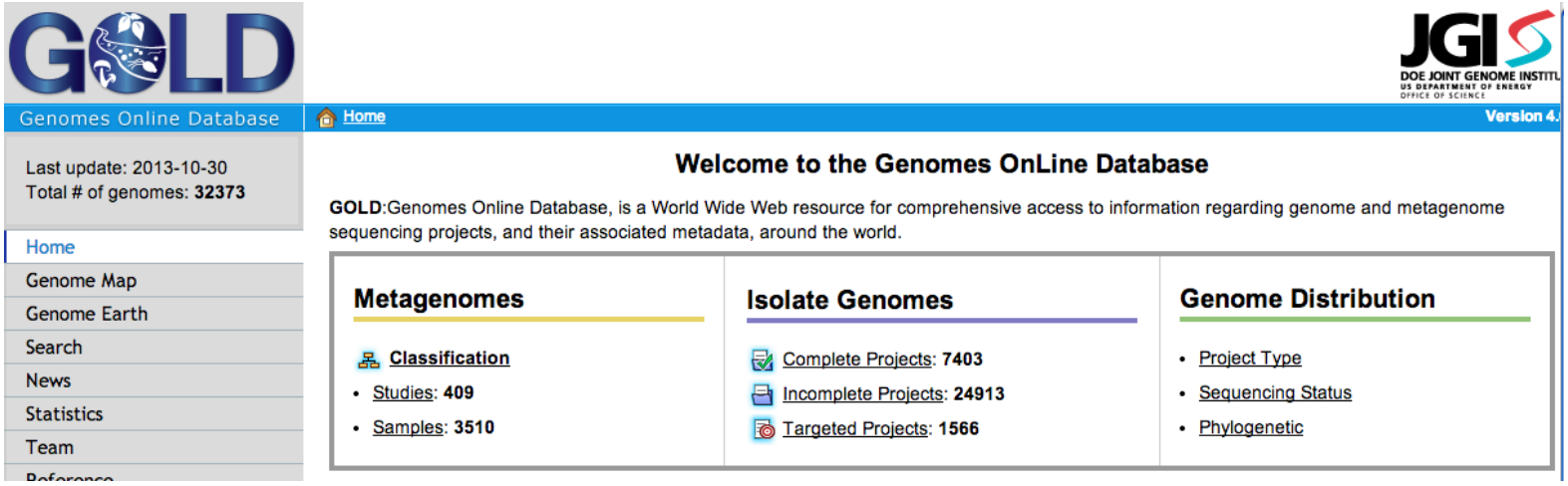
Variation du nombre de copies des ET : source de diversité génomique

Et où intervient la bioinformatique dans ces projets ?

- **Aux différentes étapes du séquençage des génomes** :
  - Lecture des séquences à la sortie des séquenceurs : qualité, adaptateur etc.
  - Assemblage des séquences à partir des fragments séquencés ou mapping
  - Annotation des génomes
- **Pour exploiter les données génomiques** : définition de marqueurs (SSR, SNP), annotation de séquences, recherche de gènes, phylogénie, génomique comparative
- **Création de banques de données** : compilation, gestion et exploitation des données

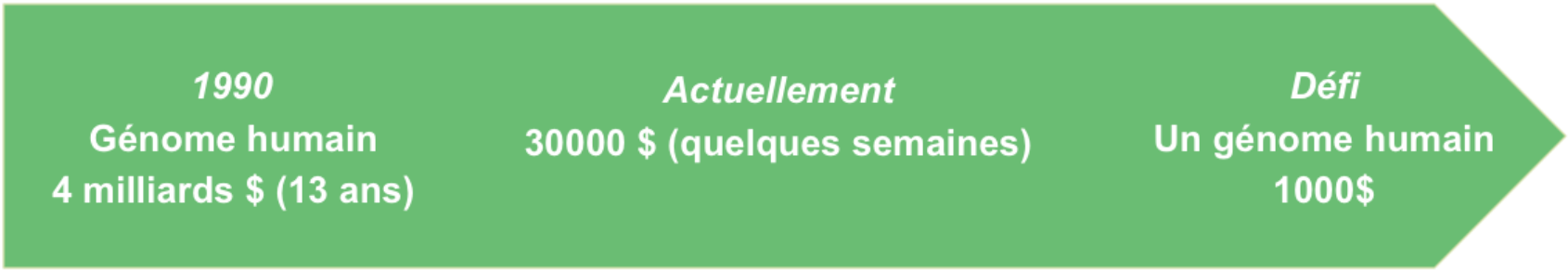
Internet, une mine d'information

- http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html
- http://www.genomesonline.org
- http://genomevolution.org/wiki/index.php
- http://plantgdb.org/
- http://www.ncbi.nlm.nih.gov/



Un peu de pratique... TP !

Comparatif des technologies NGS



De nombreuses techniques de séquençage : Roche/454, illumina/Solexa. Comment choisir ? Quelles sont les différences ?

Comparatif des technologies NGS

Platform	454 GS Junior	IonTorrent PGM	Illumina MiSeq
Instrument Cost:	$108,000	$80,490	$125,000
Sample Prep:	Emulsion PCR	Emulsion PCR	On-instrument
Run Time:	4h	3h	27h
Cost per Run:	$1,100	$425 (316 chip)	$750
Throughput/run:	71-72 Mbp	260-304 Mbp	1,653 Mbp
Avg. Read Length:	522 bp	123 bp	2 x 150 bp
Reads Aligned:	99%	90%	99%

http://massgenomics.org/2012/04/comparison-of-benchtop-sequencers.html

Les différentes étapes du projet de séquençage

Préparation de l'échantillon

- **Extraire l'ADN et le découper en petits fragments** : enzyme de restriction etc.
- **Ajouter des adaptateurs aux extrémités des fragments** : les adaptateurs incluent les amorces PCR si nécéssaire
- **Amplifier l'ADN selon les technologies**

ADN extrait prêt à être envoyé à une boîte de séquençage.

Les différentes étapes du projet de séquençage

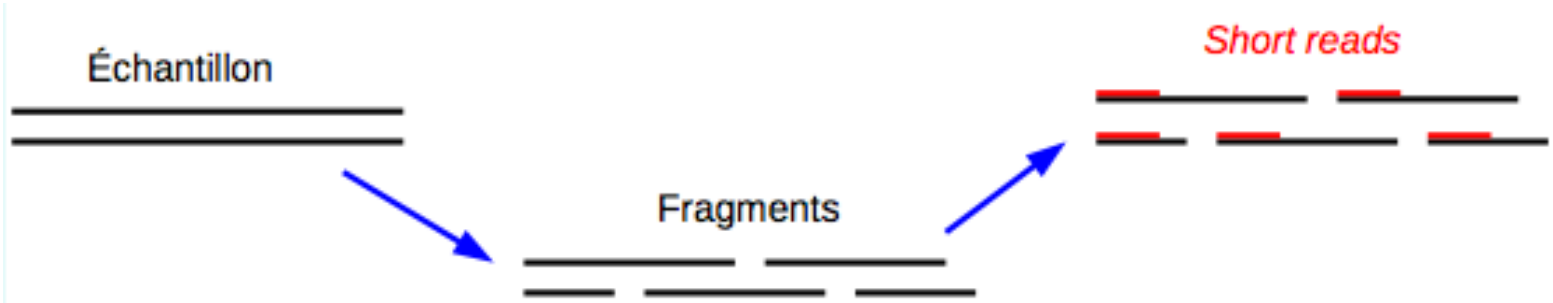
Lors du séquençage

- On effectue un nombre prédéfini de cycles de séquençage
- **Les fragments ne sont pas lus en entier**, seules les extrémités sont séquencés
- Les fragments d'ADN séquencés ont **tous une longueur identique de bases**

Les différentes étapes du projet de séquençage

A l'issue du séquençage

- Seul les extrémités des fragments d'ADN sont séquencés

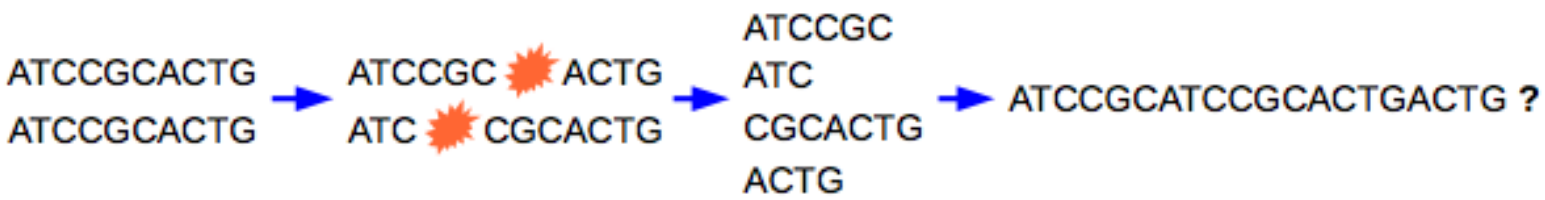


- Chaque séquence est présente en plusieurs exemplaires dans l'échantillon en recoupant les reads, il est possible d'obtenir la séquence complète

Les différentes étapes du projet de séquençage

Analyse bioinformatique

- A l'issue du séquençage.
  Ex : un run illumina pair-end 200 millions de séquences de 75 pb
- Pour une même séquence de gène, plusieurs fragments sont obtenus dont seuls les extrémités sont séquencées
- Il faut assembler les reads pour obtenir la séquence de l'échantillon
- Les fragments peuvent se recouper en partie car la molécule d'ADN n'est pas nécessairement coupée au même endroit



- Il est possible d'obtenir la séquence complète

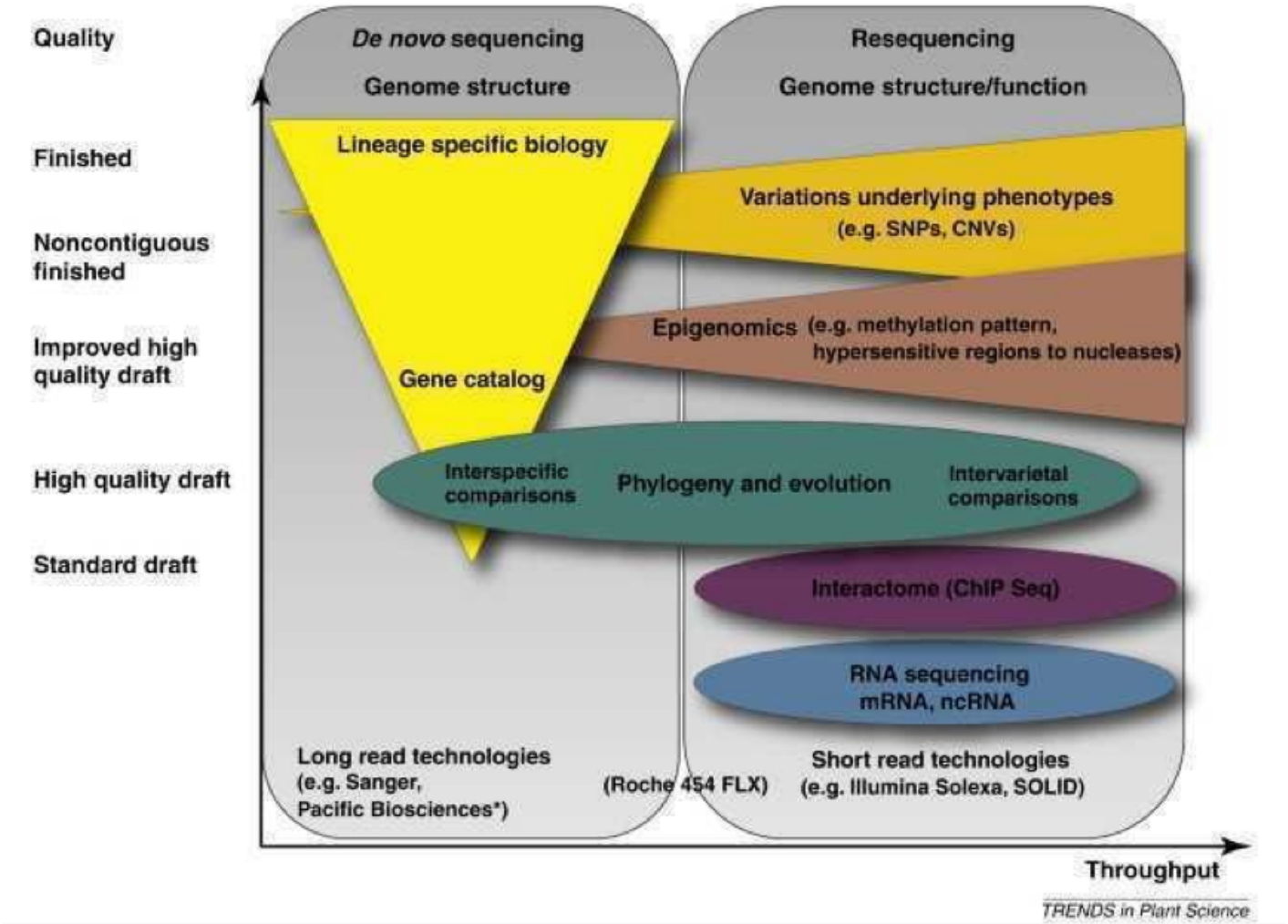
Les différentes étapes du projet de séquençage

Analyse bioinformatique : 2 approches

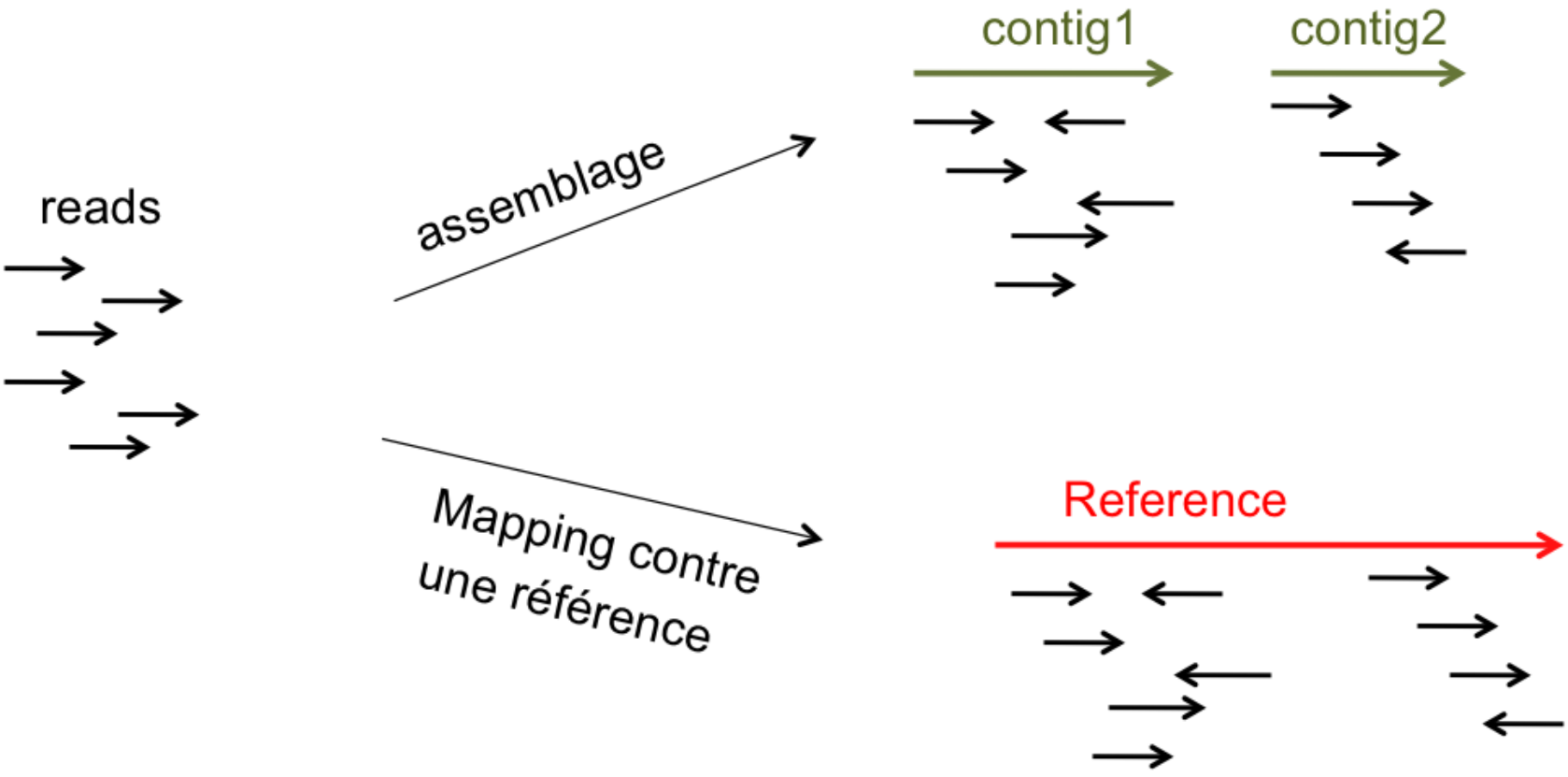
- Assemblage *de novo* : on cherche à assembler des fragments pour obtenir la séquence originelle
- Alignement / mapping avec des séquences existantes quand on a génome déjà séquencé ou un transcriptome de référence

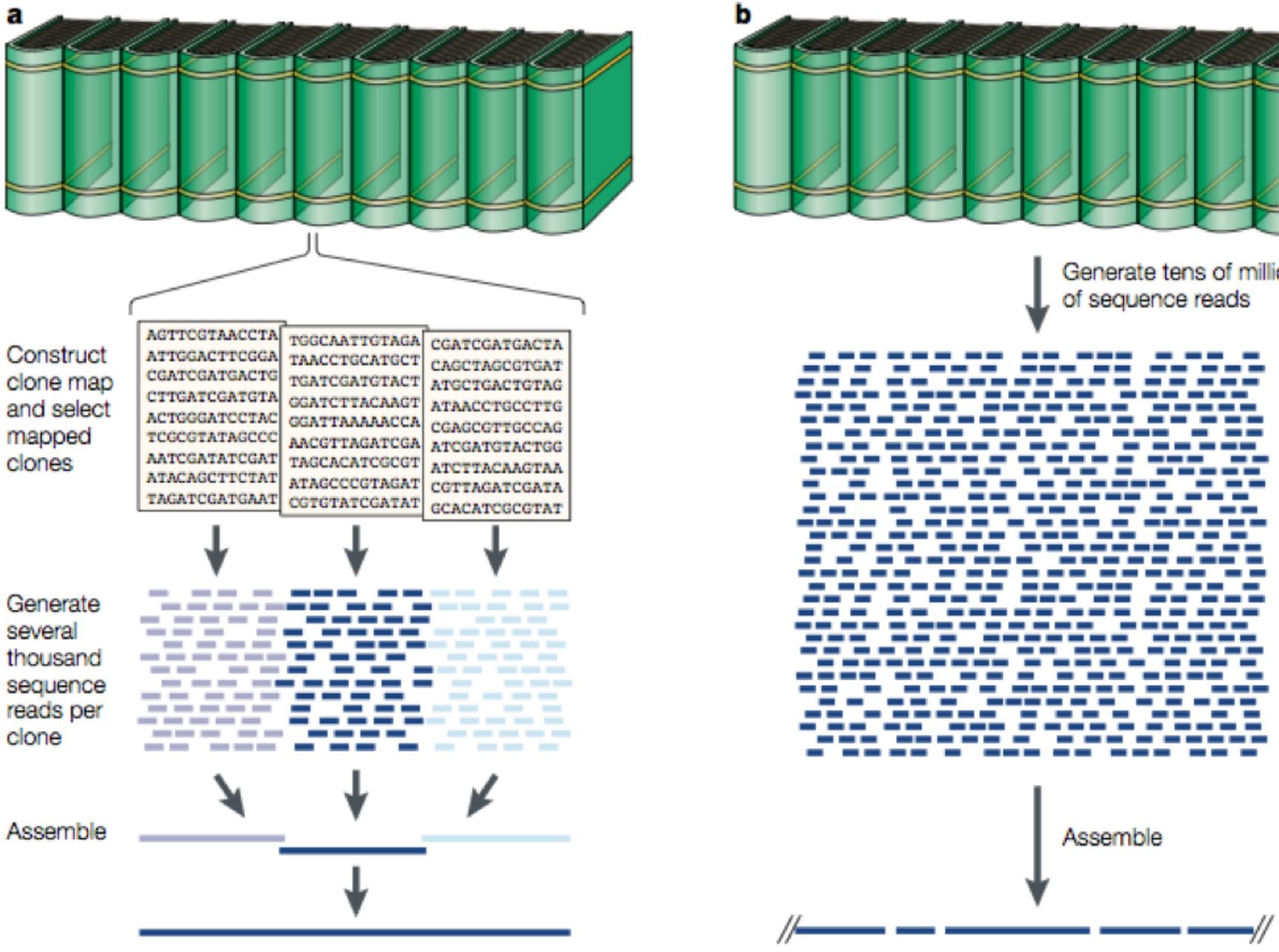
Etapes purement bioinformatiques.

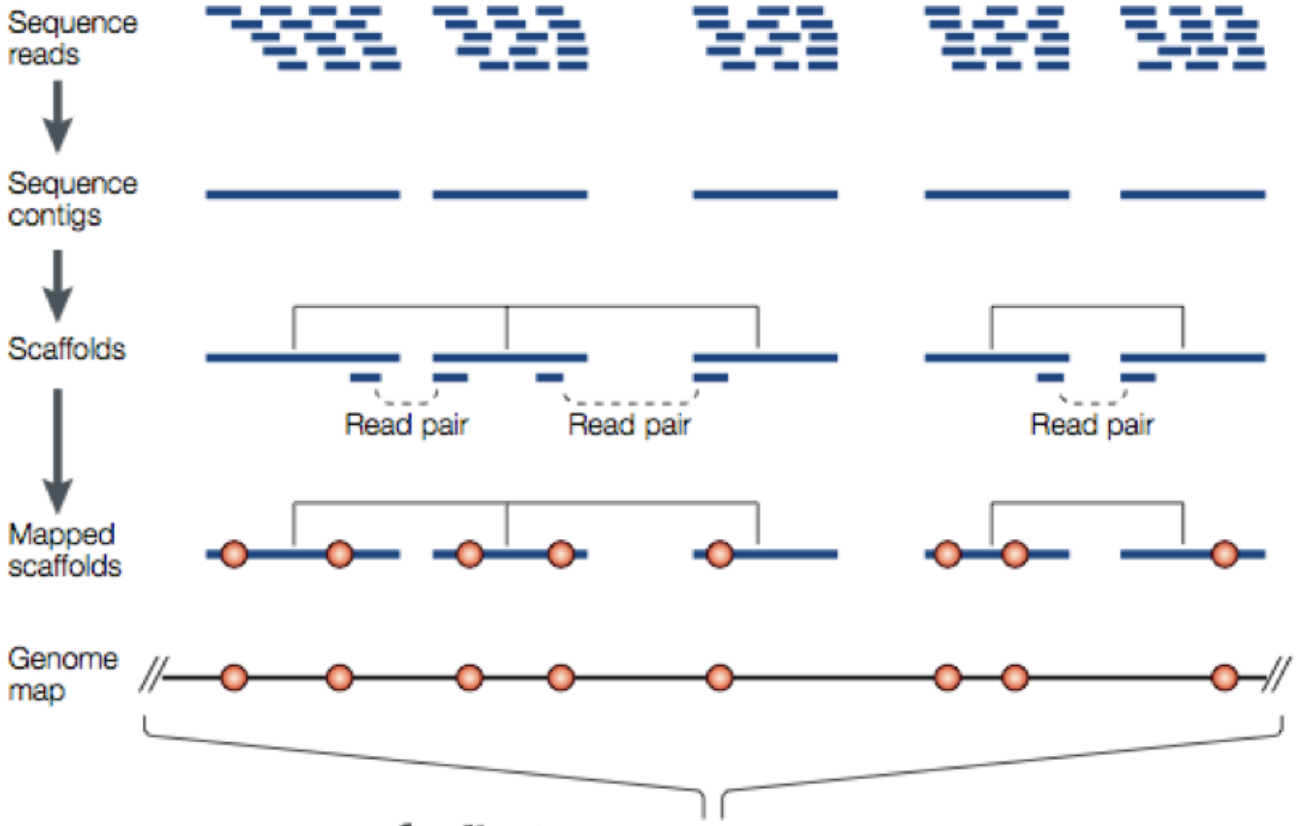
Quelles technologies utilisées pour mon projet de recherche ?

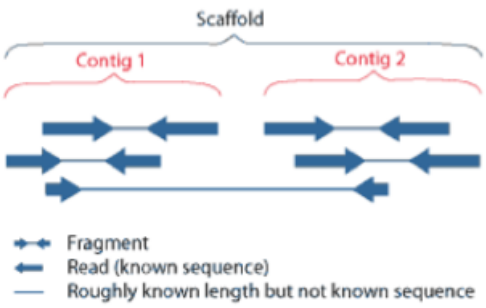


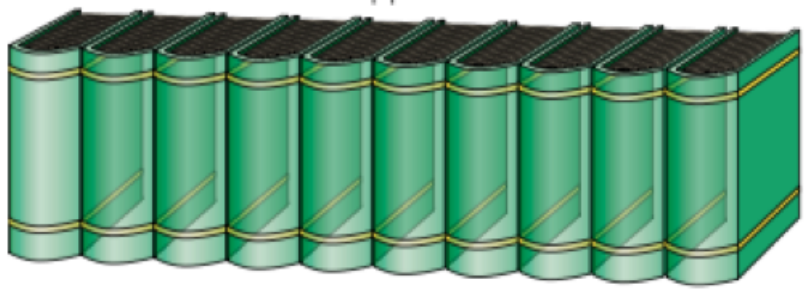

PHASE TWO • INTERPRETATION
SHERMAN
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME
Focus on Quality

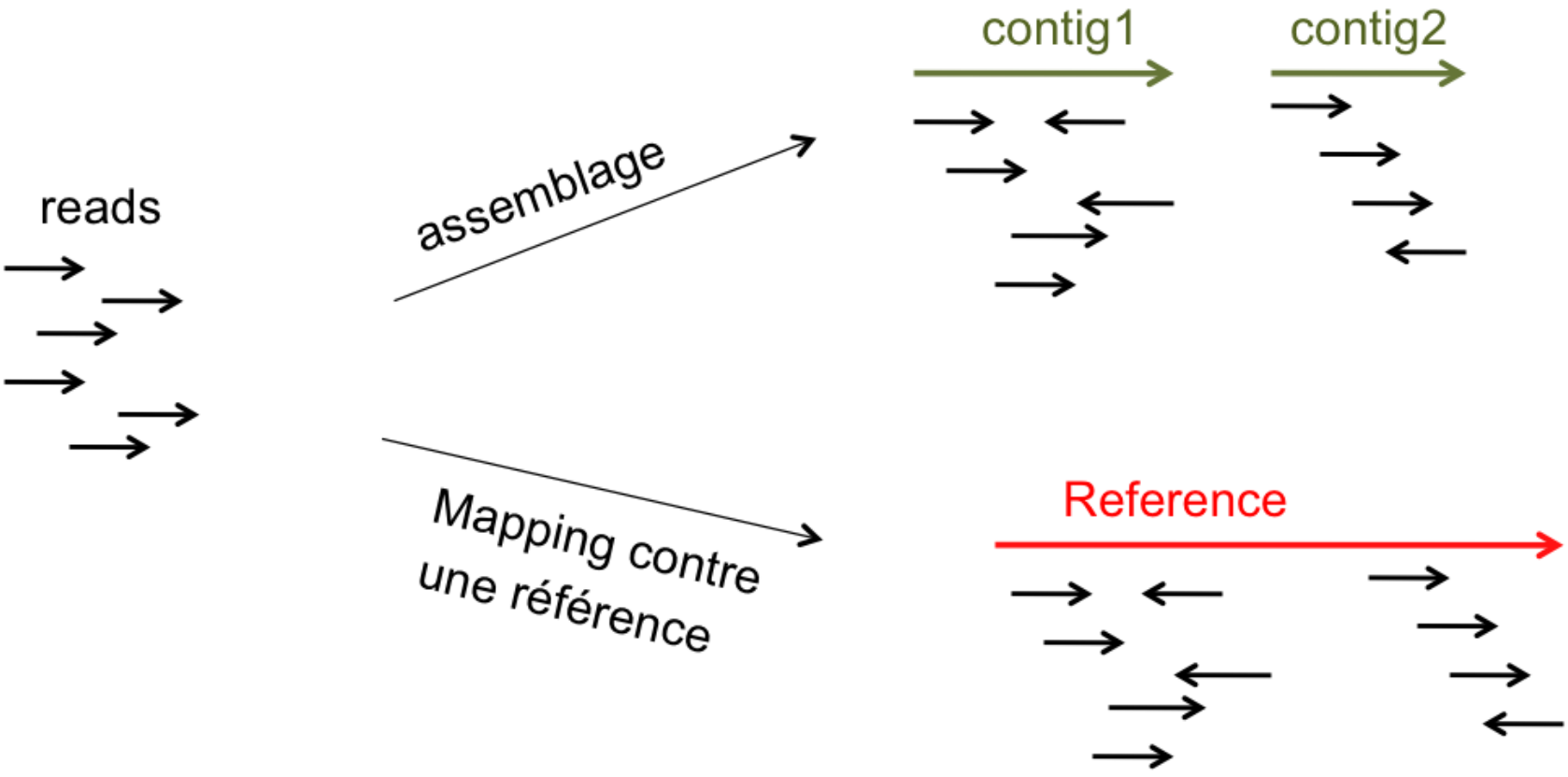

reads
assemblage
contig1
contig2
Mapping contre
une référence
Reference


a
b
Construct clone map and select mapped clones
AGTTCGTAACCTA
ATTGGACTTCGGA
CGATCGATGACTG
CTTGATCGATGTA
ACTGGGATCCTAC
TCGCGTATAGCCC
AATCGATATCGAT
ATACAGCTTCTAT
TAGATCGATGAAT
TGGCAATTGTAGA
TAACCTGCATGCT
TGATCGATGTACT
GGATCTTACAAGT
GGATTAAAAACCA
AACGTTAGATCGA
TAGCACATCGCGT
ATAGCCCGTAGAT
CGTGTATCGATAT
CGATCGATGACTA
CAGCTAGCGTGAT
ATGCTGACTGTAG
ATAACCTGCCTTG
CGAGCGTTGCCAG
ATCGATGTACTGG
ATCTTACAAGTAA
CGTTAGATCGATA
GCACATCGCGTAT
Generate several thousand sequence reads per clone
Assemble
Generate tens of millions of sequence reads
Assemble


Sequence reads
Sequence contigs
Scaffolds
Read pair
Read pair
Read pair
Mapped scaffolds
Genome map


Scaffold
Contig 1
Contig 2
Fragment
Read (known sequence)
Roughly known length but not known sequence




reads
assemblage
contig1
contig2
Mapping contre
une référence
Reference

Analyses bioinformatiques

Logiciel de qualité, assemblage, mapping

Quelques questions :

- Quelles ressources de calcul et de stockage ?
- Quels programmes utilisés ?
  - Très nombreux, en évolution permanente en parallèle des technologies de séquençage
  - Ne pas les croire systématiquement !

Importance de l'étape de test des logiciels avec des jeux de données

CAACTAGCAGCTAACCAGCAGAGAAAAATTAATGAGATTGAAAACAGCTGCAAGCACAAGAAAATTCTGTACAAAGCTCCATCC
AAGAATTGAGAGAAGAACAAAGTCAGCTCACTAGCTCCATCACTGCACTAAATAATAAAATTGATAAGCTAATGGAAGAGCCAAA
GGAAGATCAATCTGGACCTTCATCTTCAAAACCTTCAACCTAGACTGACCTTAGACTAAACCTAACCTAAGCTTAGACTGAAACC
CATGTTTTCAAAATATAATCTTGTACTAAGCTGAGTTTCATGATTATTTGCCACACTTGTATGTACAGATACCAATTGATAACAATTG
CAATTCTCTGATTTATCAAATGTACTTTATGCAATGAATGATTACCTTTCTTGAAACATATGTCTTCATTGGTTTATATACTGCTCTG
ATACATATTGCATCTGATACAAATCAAAATCTGTAATGGTATCAAAATTCTTAATGTTGGTATCAGATATTATTTTTAACTTCCTTCCT
TTTGTTGATGACAAAAAGGGGGAGAAATATAAAAAAGAATTCTCATCTGAGAAATGAGAAATATATGAAAAAGAATTCTAATCTGA
GAAATAAGAAATGTCAAAAGAATTCAGAATACAAATCTCTGCACTAAAGCTAATATATAGGGGGAGATTTAAATTATGATAAAGAAC
AAGAACTCTGCATAATTCGAATGTTTGTAATTGATAGGGGGAGACCTTTCTCCATAATCTGTTGAAATTATGCATATATCTCTGAAC
GTCATATTTGCTCTGAACTTCATTATAGAATTCTAATAATTGCTGTGAAATTTATTGTAAAATTTATTTTAACTCAATGGTTTTGTCAT
CATAAAAAAATGGGGAGATTGTCAACCCTAAAGGATGAATTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNTGTCTGAAAACCAACTGGGTTATTTGTGATCGCAGAAAACAAAAGGTTTTCGATTGGTTGCCT
GAGAAAACCAACTGAATTTTGATTGTGAGCCCGGAAAACAATCAGGCTGTAATCTGCTGGGTTATAGTGAAATCTCAAGCTAGG
CTTGAGGAGTGGACGTAGGTGCTGGGAGTGCATCGAACCACTATAAATCTTGGTGTTTGTGATTGTGCTTCTTCTCTTCCTTTC
CTCTGCATATTCCTGACATTTCTCATAACTTTATTCACTGTCATTTGTACATTCTTATTTCCGCTGCTCAATCTTTAAAATAAAAGT
AACTCATACTCTCTTCACGTTTTAACTTTAACTATTTTTTAAAGACCCCAATTCACCCCCCCTCTTGGGTTGCACCTCTGGGCAA
CAATGAGCAATACCAATATAGTATTGAACGATACATATATACACTCTATTATAAGCTAAAATGAATGTGGAGAGTGATAAACATATTT
CCAATACAAAGGCAAAATTATAGTTTTAGCCCATGAGCATTCTTGGTTGATGATGATCATTGTTGTTATATCTCCTGGATTCATGG
TACATTCCGTTGCTTTACCTTTCTATTGGCCAAAATCTCCTCCATTTGTTCTATAATTTTGGACTAAAAATGAGTTAGATTTAATCTG
ACCGAATCTATTTCCATATCTGAAGTGTATCAGATTCGAACACAAATTTTGATATCCGAATTGATGTGGATCCGAATTTGGATTTTA
AATACTTTTCGAATCCAAATCTAGATATCCAGTAAAAAAAAATTCAAAAATAAACTGCATAACCTAGGATTCCAACTAGAGATATATA
GTTTGAATGGTAATGTCTTAAACACCAAGATACAATTATATTCCTATCTTATTAAATGAGTTTATTTTACTTGTTTATATGTATTCAGG
TCATTAAATATCATATAATGTATAGCATGATATTTCATATTTTAATCACTATTTTTTATAAAATACTTCTCTTTGATAAAAATGAAAATGA
ACATACTTTATTTTATTATTTATAATATGTTTGAAACTTTATTGTTAAAATATTTAAAAATTAACTTCTATTGTTCTTATATGCATATATAAT
TTTGCTTACAAGCTTTATAAGAATATATAATTCACAATTATTTTTATTTATGCTATAATTAAAAGTAATAATAATAATTGATACAAACATG
CTAGTATTTTTATGTTGTATTCTTATCTTAAATGATTTTTATTTAATATTTTAATAAAAATCTGGTCAAACCAATGACCTAATCCTTGAC
AAGGTCAGTCTTCATGTCAGTTTAATAACCATATCTTAGATCCTTGCTTACTTGCGTATCGTGTGCAATTTGCATAGATAAAAATAA
TATCCGATTTATATTTGTAGTTATTTAAAACAAATATGCATATGCTTAATATCCGATCTATATCCGTATCTATTTAGAACACAATGCAAA
CTAATTTGGTATTCATTTTATATCCATATGGATCTATATTTGTTAAAAAATATGGATACAAATATGGATGTACCAGTATTCGATCAATAT
CCAATTTGTTTTTACTCCATTAACATCACGATAGTGGCAGATATCATGGGACTTATACTTTTGCTTGCAAGCTCTTCTTTTTAGTGG
AGGTCTTGCCAATGCTTTAGTTGGGACATCATAGGATTCTATGGATGGTGGGGCTTATATTTTTTTCCTTGGACCCTGCCTAATA

Annotation structurale

Positionner les éléments génétiques sur la séquence génomique
.... De manière précise, complète et exhaustive

Annotation structurale

Positionner les éléments génétiques sur la séquence génomique .... De manière précise, complète et exhaustive

En pratique, le plus souvent, positionner les gènes et leurs produits : transcrits, protéines ..

mais aussi – quelquefois – d'autres objets, comme les éléments transposables, les motifs de régulation, les domaines, etc...

BLAST
Glimmer Exonerate
RepeatMasker
consed
GeneMarkHMM
FGENESH

BLAT
RepeatScout
LTR_STRUC
RepeatFinder
Cap3
Dotter

ARTEMIS
ACT
APPOLO
GBROWSE

Annotation structurale

Principe

Différentes méthodes selon les données à disposition

- Méthode expérimentale : utilisation de transcrits (cDNA) complet et provenant du même organisme
- Méthodes comparatives (extrinsèques)
  - Traduction de la séquence génomique en protéine et comparaison aux séquences de protéines connues (banques de données)
  - Comparaison aux séquences d'ESTs disponibles
  - Comparaison aux séquences génomiques provenant d'espèces proches
- Méthodes *ab initio* (intrinsèques) : Recherche des particularités communes à tous les gènes de notre génome, puis détection sur le génome
- Méthodes intégratives : pourquoi ne pas combiner ces approches ?

Annotation structurale : comment comparer ?

Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?

Existe-t-il des séquences homologues à la mienne parmi toutes les séquences connues ?

Banque UniProt, 12 millions de séquences, 350 AA/seq

Annotation structurale : comment comparer ?

Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?

Existe-t-il des séquences homologues à la mienne parmi toutes les séquences connues ?

Banque UniProt, 12 millions de séquences, 350 AA/seq
Smith & Waterman : 0.035 s x 12 millions 118 heures 5 jours ! !

Annotation structurale

Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?

Solution 1

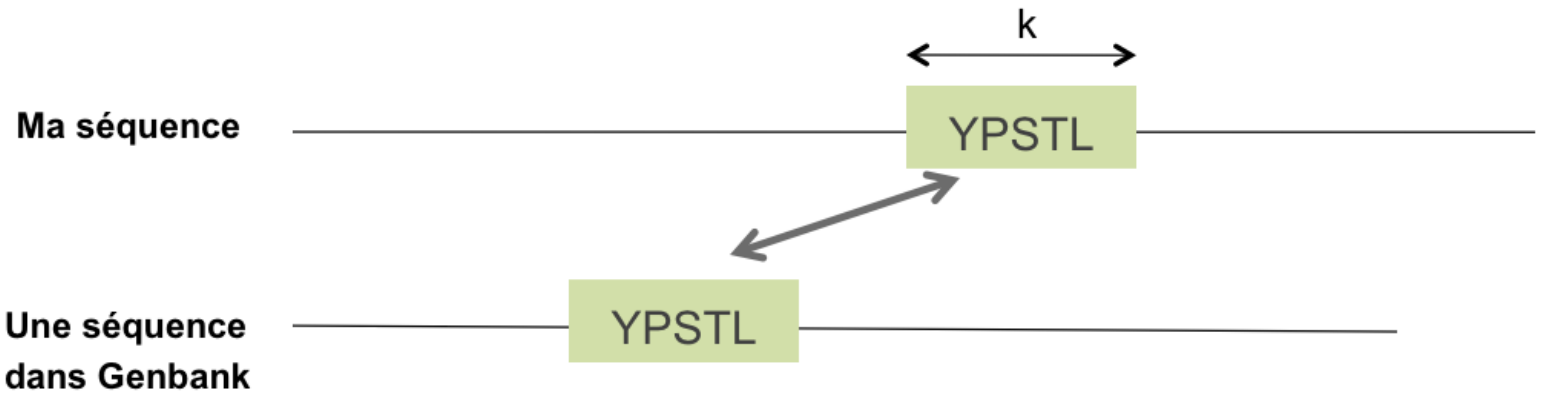
- Comparer votre séquence (600 aa) avec chaque séquence de la banque (Genbank : 85 millions de séquences)
- Avantage : La séquence la plus similaire
- Inconvénient : Temps de recherche
- Algorithme : Smith-Waterman
- Alignement global
- Programmation dynamique
- EXACT

Solution 2

- Faire une pré-sélection sur les séquences puis aligner exactement avec SW
- Avantage : Rapide et efficace
- Inconvénient : Risque de passer à côté de la perle !
- Algorithme : Smith-Waterman
- Alignement local
- BLAST
- Heuristique

Annotation structurale : comment comparer ?

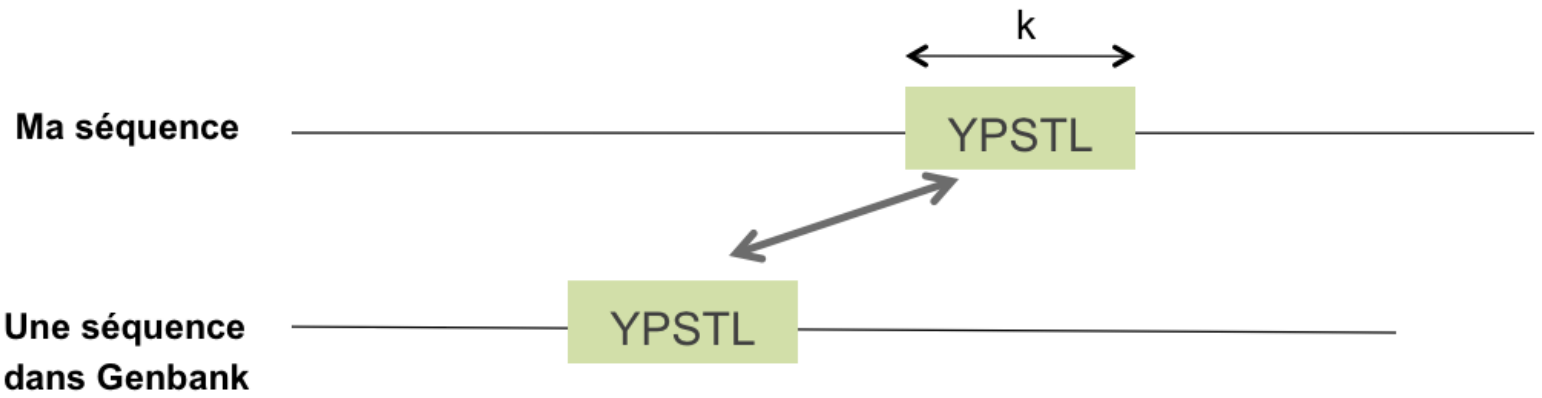
Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?



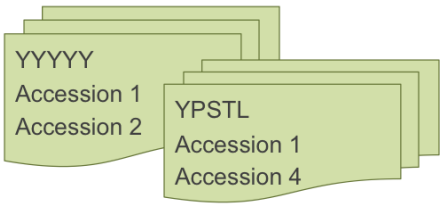
Ne retenir que les séquences partageant au moins
un mot de longueur k avec ma séquence
Pourquoi cette sélection est si rapide ?

Annotation structurale : comment comparer ?

Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?



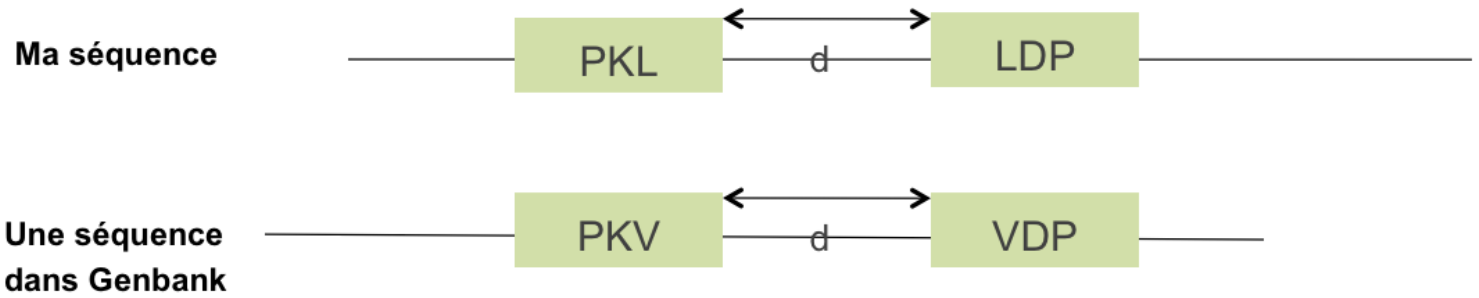
Ne retenir que les séquences partageant au moins un mot de longueur k avec ma séquence



BLAST indexe les séquences et détermine, pour tous les mots de longueur k, la liste des séquences qui contiennent ce mot

Annotation structurale : comment comparer ?

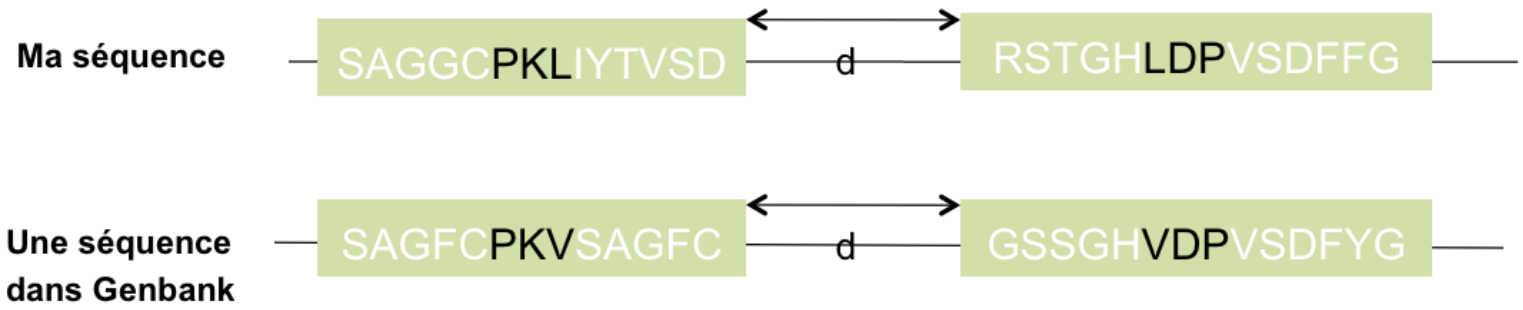
Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?



Trouver 2 paires de mots voisins, s >= à 11 et à égale distance avec d<40 dans les 2 séquences

Annotation structurale : comment comparer ?

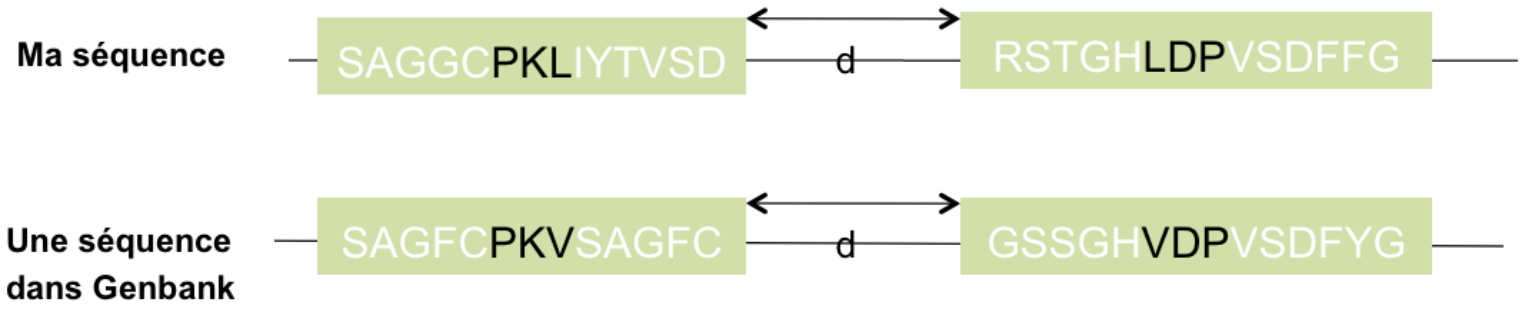
Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?



Vérifier que l'on peut étendre ces amorces pour obtenir des alignements sans gaps de score s>S

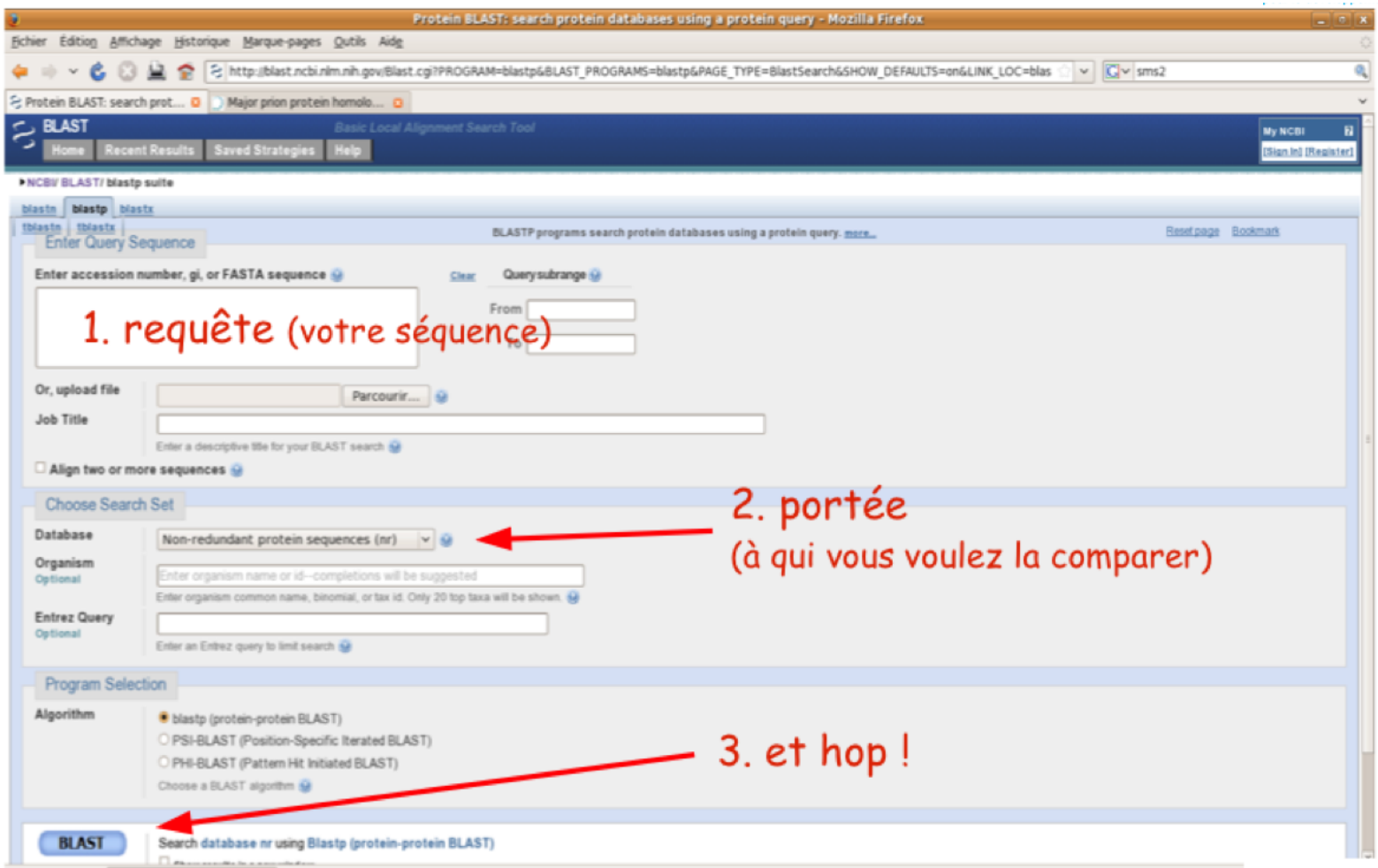
Annotation structurale : comment comparer ?

Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?

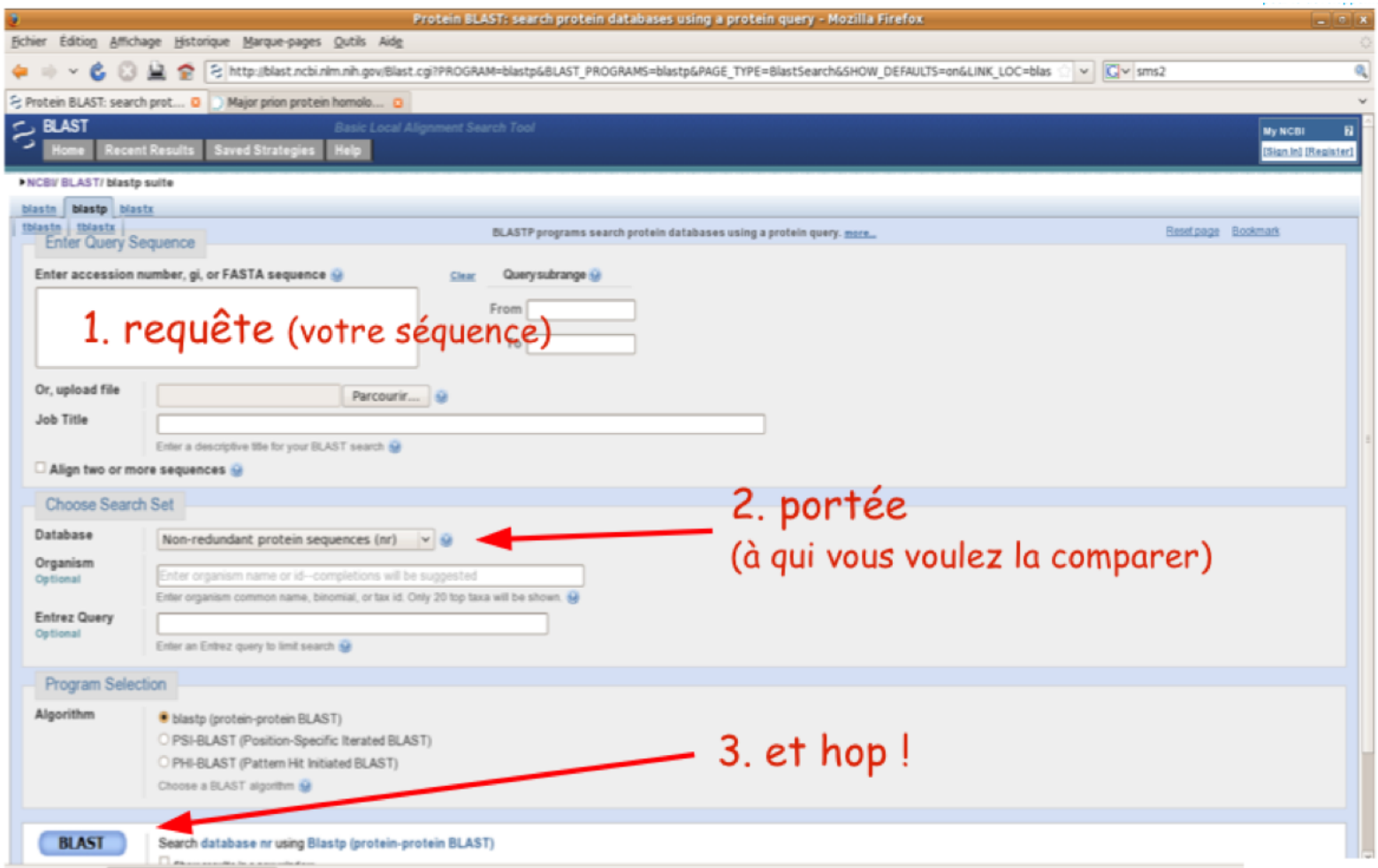



```
SAGGCPKLIYTVSD
|| |  | ||   ||  |
SAGFCPKVVYTLSE
```

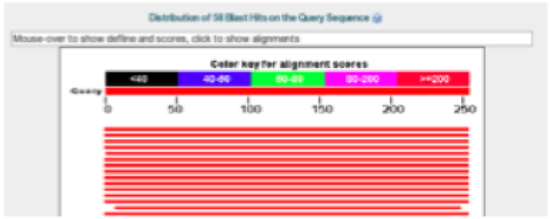
Annotation structurale : Blast

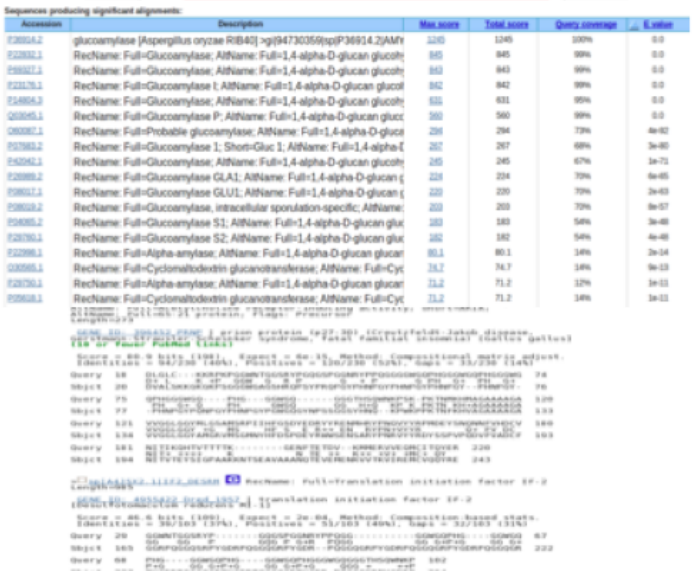


Annotation structurale : Blast



Annotation structurale : Blast



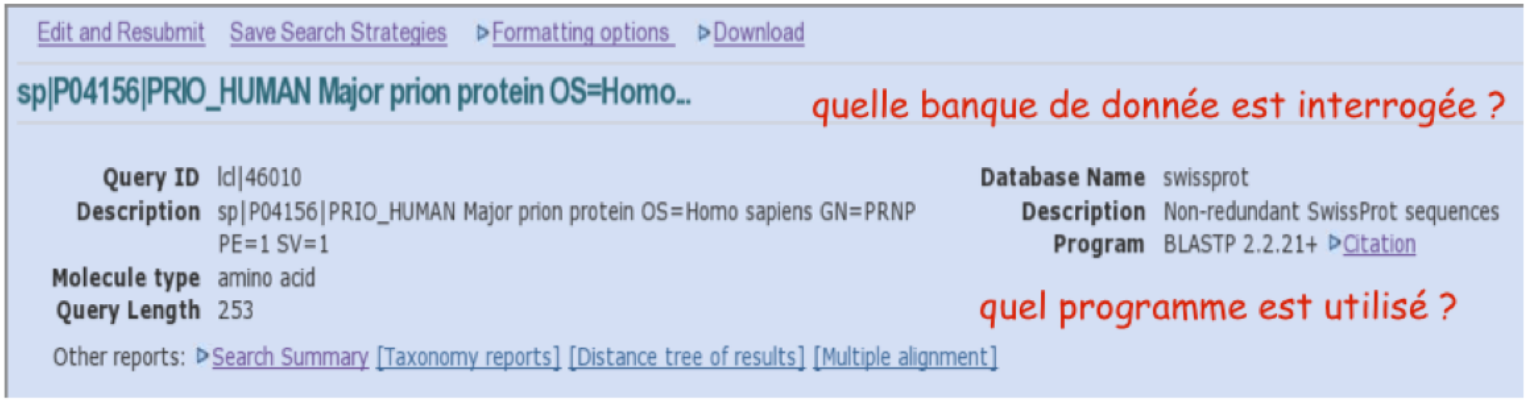


1. récapitulatif de la requête
2. représentation graphique des résultats
3. résumé des résultats
4. les alignements

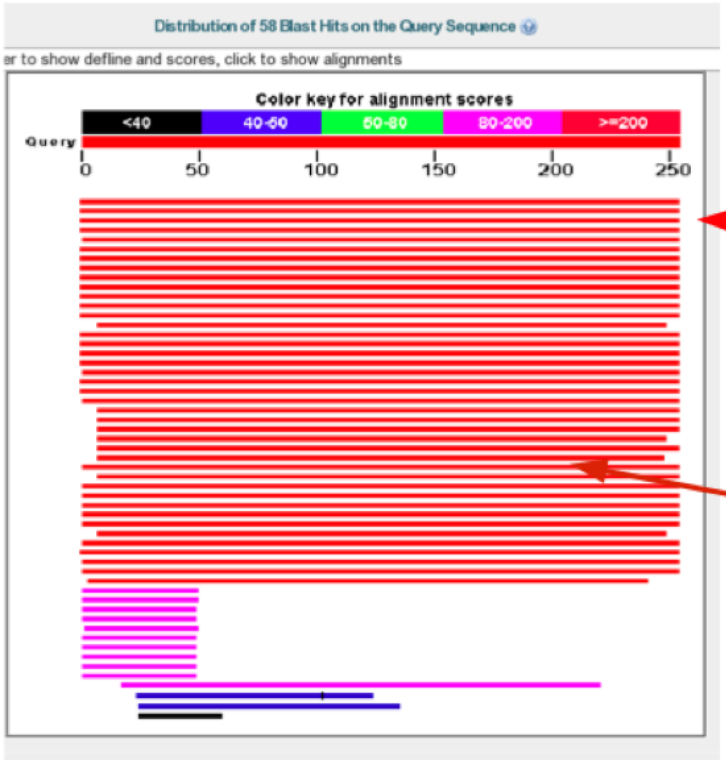
Annotation structurale : Blast ?

1. récapitulatif de la requête

quelle séquence a été soumise ("query") ;
identifiant, longueur, type



Annotation structurale : Blast



2. représentation graphique des résultats

ce trait représente
la séquence soumise (long. 253 AA)

chaque trait de couleur représente
un alignement entre la séquence de
départ et une séquence de la
banque de donnée sélectionnée
couleur → score
longueur → taille de l'alignement

= HSP ("high scoring pair")

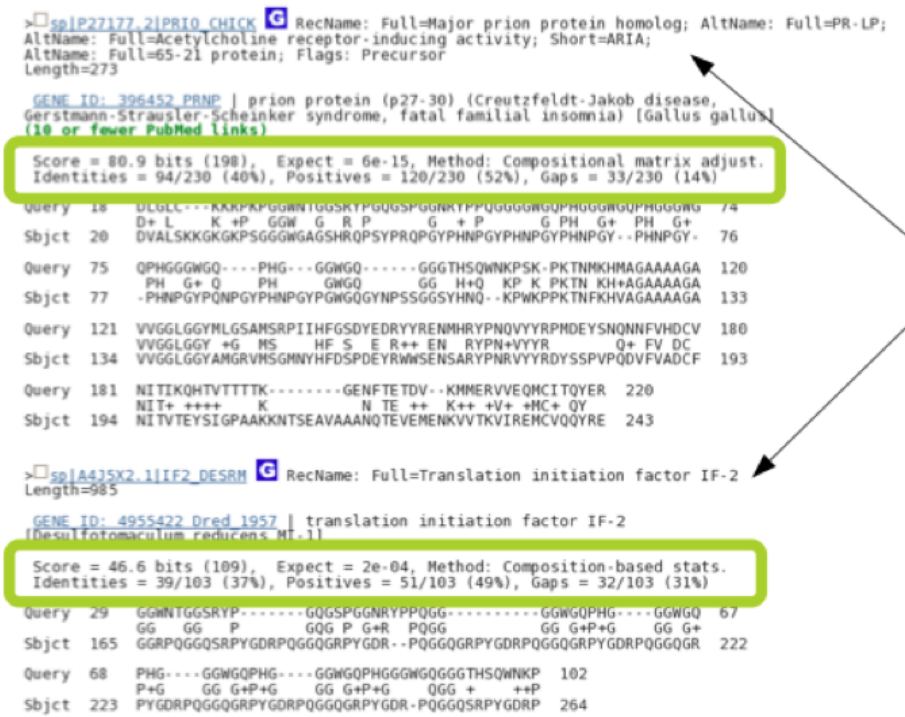
Annotation structurale : Blast

identifiant descriptif score couverture E-value

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
P36914.2	glucoamylase [Aspergillus oryzae RIB40] >gi|94730359|sp|P36914.2|AMY	1245	1245	100%	0.0
P22832.1	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucohy	845	845	99%	0.0
P69327.1	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucohy	843	843	99%	0.0
P23176.1	RecName: Full=Glucoamylase I; AltName: Full=1,4-alpha-D-glucan glucoh	842	842	99%	0.0
P14804.3	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucohy	631	631	95%	0.0
Q03045.1	RecName: Full=Glucoamylase P; AltName: Full=1,4-alpha-D-glucan glucc	560	560	99%	0.0
O60087.1	RecName: Full=Probable glucoamylase; AltName: Full=1,4-alpha-D-gluca	294	294	73%	4e-92
P07683.2	RecName: Full=Glucoamylase 1; Short=Gluc 1; AltName: Full=1,4-alpha-D	267	267	68%	3e-80
P42042.1	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucohy	245	245	67%	1e-71
P26989.2	RecName: Full=Glucoamylase GLA1; AltName: Full=1,4-alpha-D-glucan g	224	224	70%	6e-65
P08017.1	RecName: Full=Glucoamylase GLU1; AltName: Full=1,4-alpha-D-glucan g	220	220	70%	2e-63
P08019.2	RecName: Full=Glucoamylase, intracellular sporulation-specific; AltName:	203	203	70%	8e-57
P04065.2	RecName: Full=Glucoamylase S1; AltName: Full=1,4-alpha-D-glucan gluc	183	183	54%	3e-48
P29760.1	RecName: Full=Glucoamylase S2; AltName: Full=1,4-alpha-D-glucan gluc	182	182	54%	4e-48
P22998.1	RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucan	80.1	80.1	14%	2e-14
O30565.1	RecName: Full=Cyclomaltodextrin glucanotransferase; AltName: Full=Cyc	74.7	74.7	14%	9e-13
P29750.1	RecName: Full=Alpha-amylase; AltName: Full=1,4-alpha-D-glucan glucan	71.2	71.2	12%	1e-11
P05618.1	RecName: Full=Cyclomaltodextrin glucanotransferase; AltName: Full=Cyc	71.2	71.2	14%	1e-11

Annotation structurale : Blast



4. les alignements

query → la séquence soumise

subject → la séquence trouvée dans la bdd

alignement = outil QUANTITATIF

- scores
- Expect (ou E-value)
- % identité
- % positif
- # de gaps

Annotation structurale : Blast

Annotation structurale : Blast

cet alignement est plus réaliste ...
(1 seul événement évolutif)

```
CGATGCAGCAGCAGCATCG
||||||       |||||||
CGATGC------AGCATCG
```

.. que celui là !!
(5 évènements évolutifs)

```
CGATGCAGCAGCAGCATCG
||  ||  ||||  ||   || |
CG-TG-AGCA-CA--AT-G
```

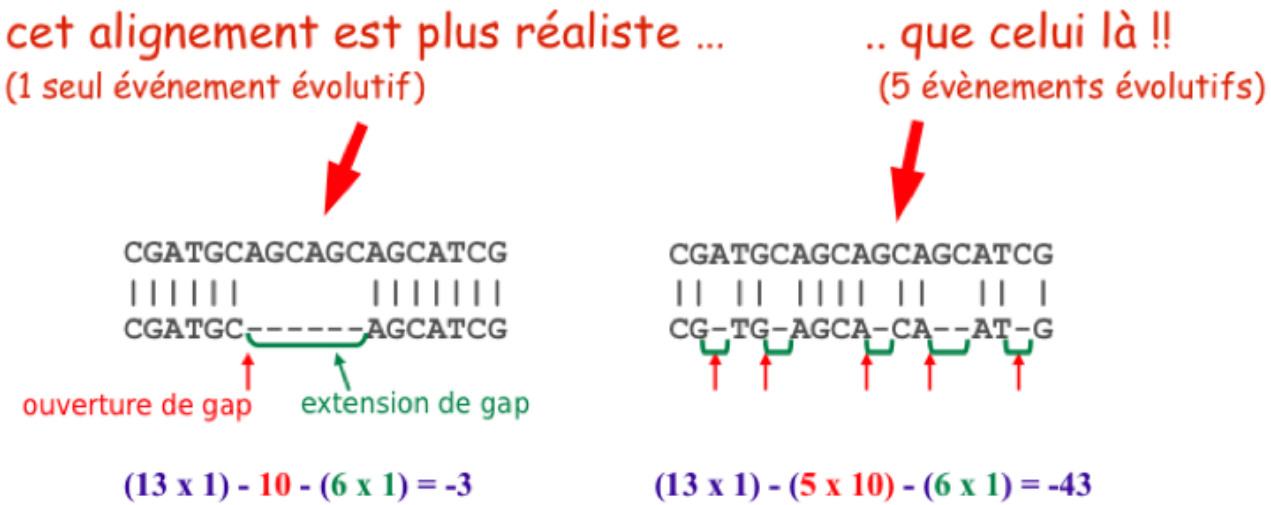
Match = +1
Gap = -1

(13 x 1) + (6 x -1) = 7

(13 x 1) + (6 x -1) = 7

- les 2 alignements ont le même score

Annotation structurale : Blast



- **insertion/délétion**: 2 paramètres
  - ouverture de gap (par ex -10)
  - extension de gap (par ex -1)

Blast : signification de la e-value

E-value

seuil de significativité statistique pour conserver un match dans les résultats.

E-value de 10

on s'attend à ce que 10 matchs similaires à celui obtenu soient trouvés simplement par hasard,

Blast : signification de la e-value

E-value

seuil de significativité statistique pour conserver un match dans les résultats.

E-value de 10

on s'attend à ce que 10 matchs similaires à celui obtenu soient trouvés simplement par hasard,

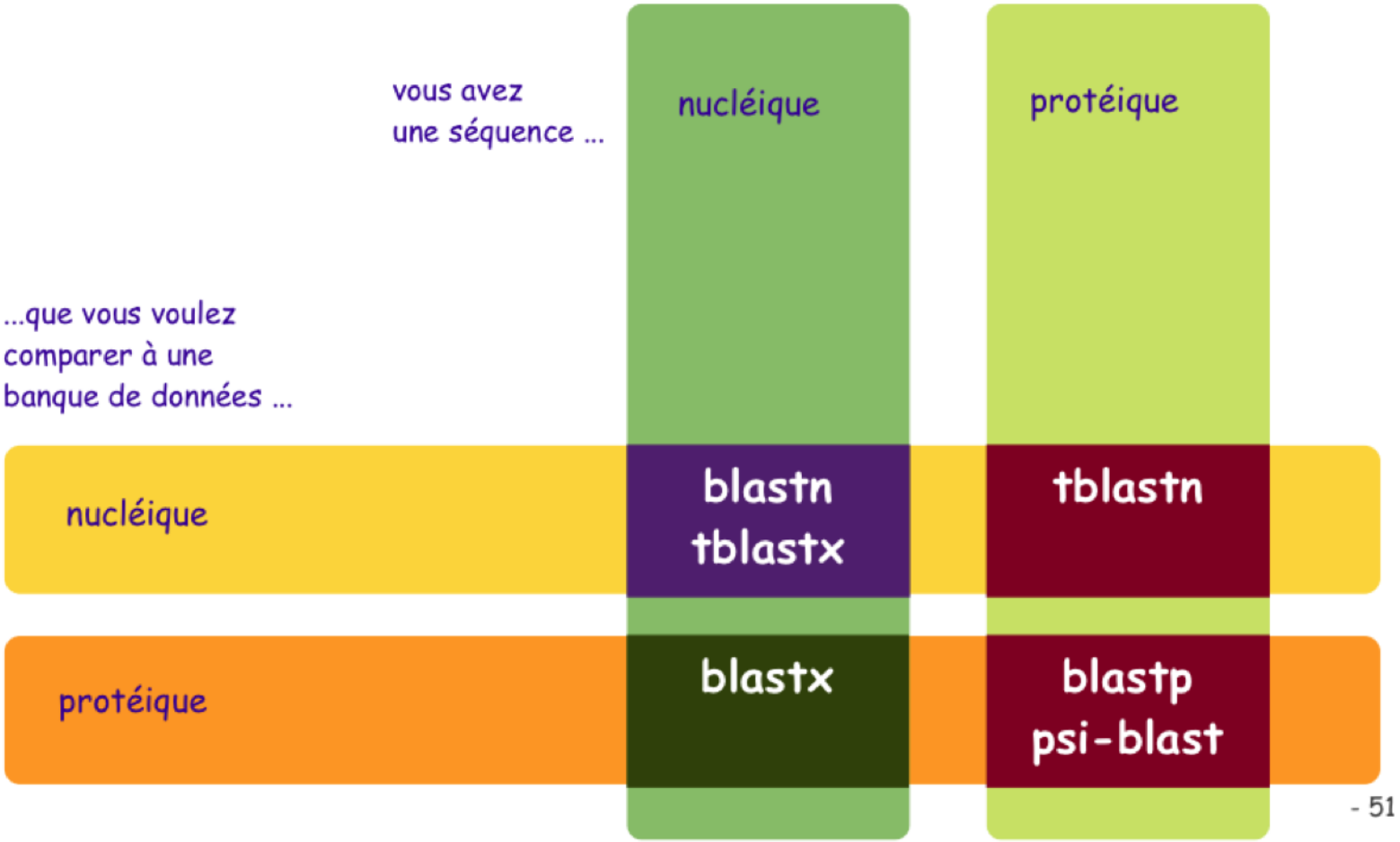
s = 46 > Evalue = 4e-4 : je m'attends à trouver en moyenne 0.0004 alignements de score 46 purement par hasard (si je blaste 2500 séquences aléatoires, j'en obtiendrai ~ 1)

s = 267 > Evalue = 1e-70: il faut que je blaste 1e70 séquences aléatoires avant de tomber au hasard sur un alignement de cette qualité ...

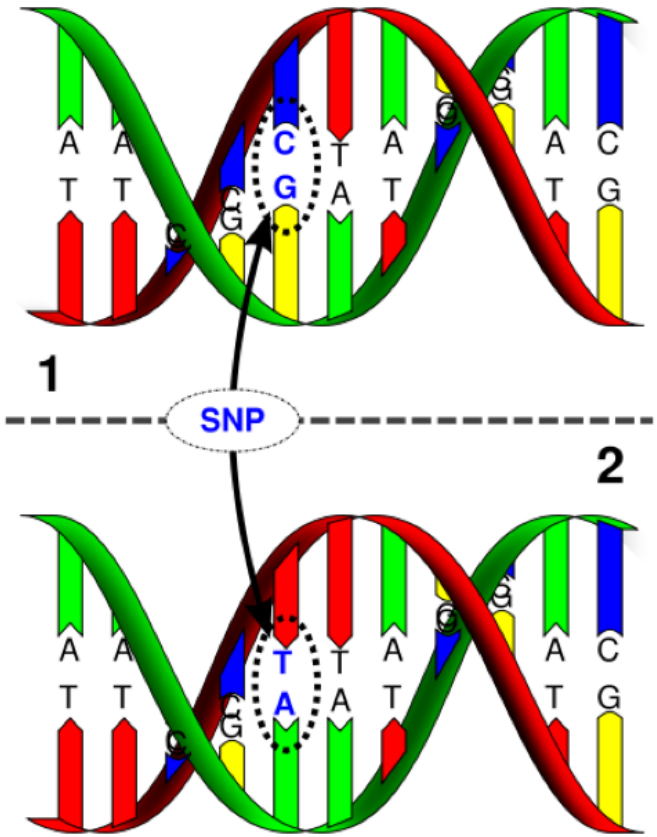
Annotation structurale : Blast

faux-positifs: on a un alignement, mais les séquences ne sont pas homologues

Annotation structurale : Blast



Après le séquençage



Next Generation Sequencing: de novo sequencing; re-sequencing and mapping

Types of variants

SNPs

Alignment	*VCF representation*		
ACGT	POS	REF	ALT
ATGT	2	C	T

Insertions

Alignment	*VCF representation*		
AC-GT	POS	REF	ALT
ACTGT	2	C	CT

Deletions

Alignment	*VCF representation*		
ACGT	POS	REF	ALT
A--T	1	ACG	A

Complex events

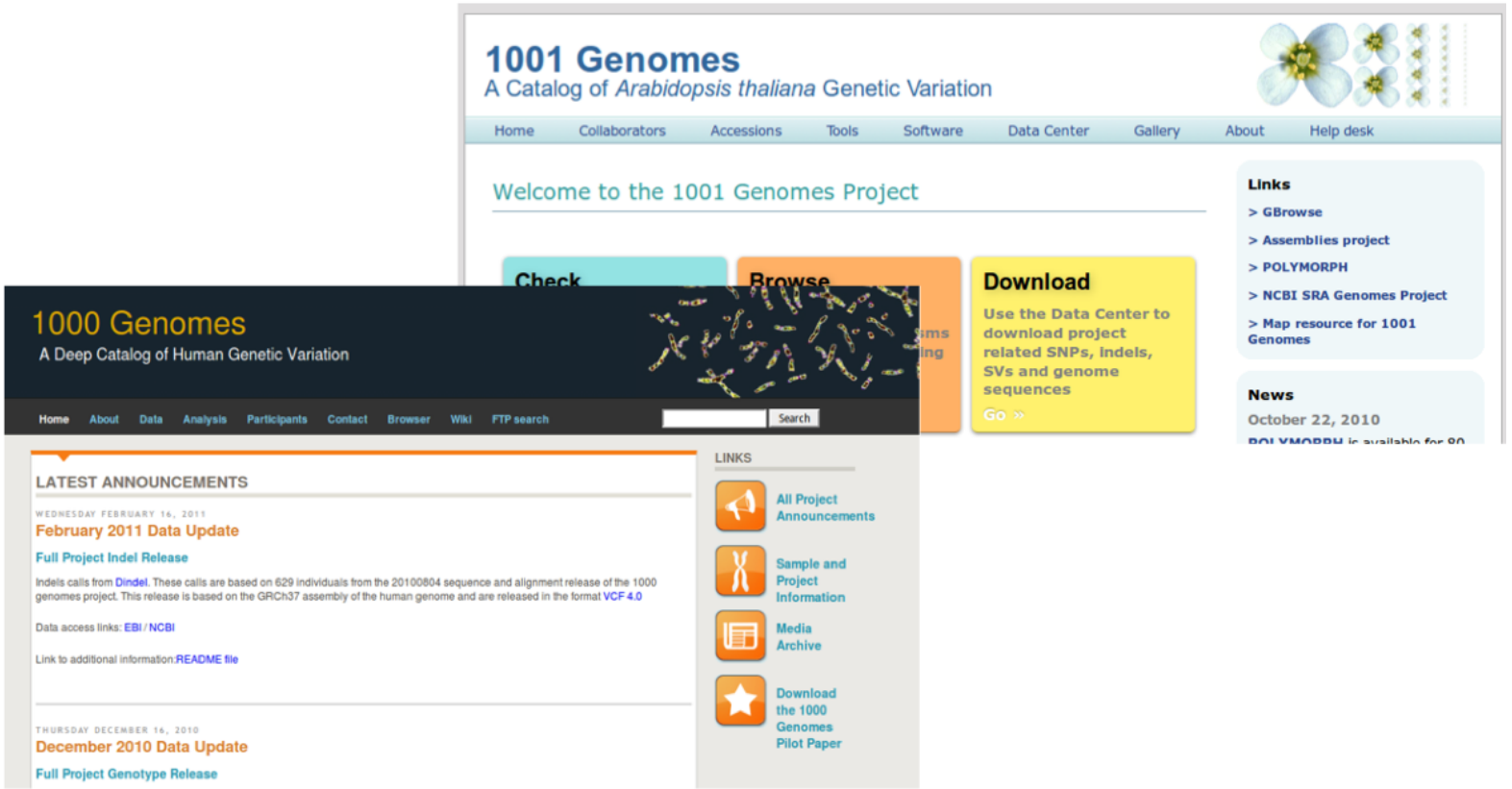
Alignment	*VCF representation*		
ACGT	POS	REF	ALT
A-TT	1	ACG	AT

Large structural variants

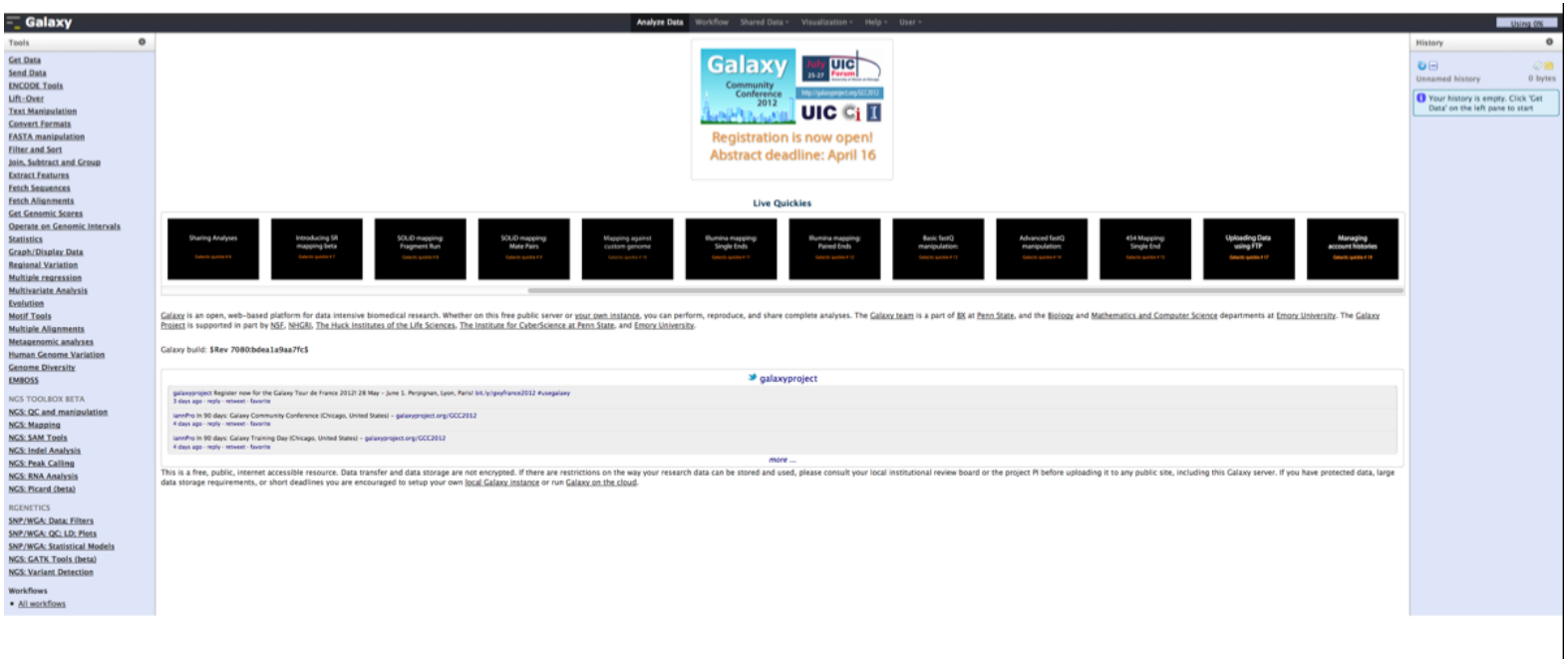
VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

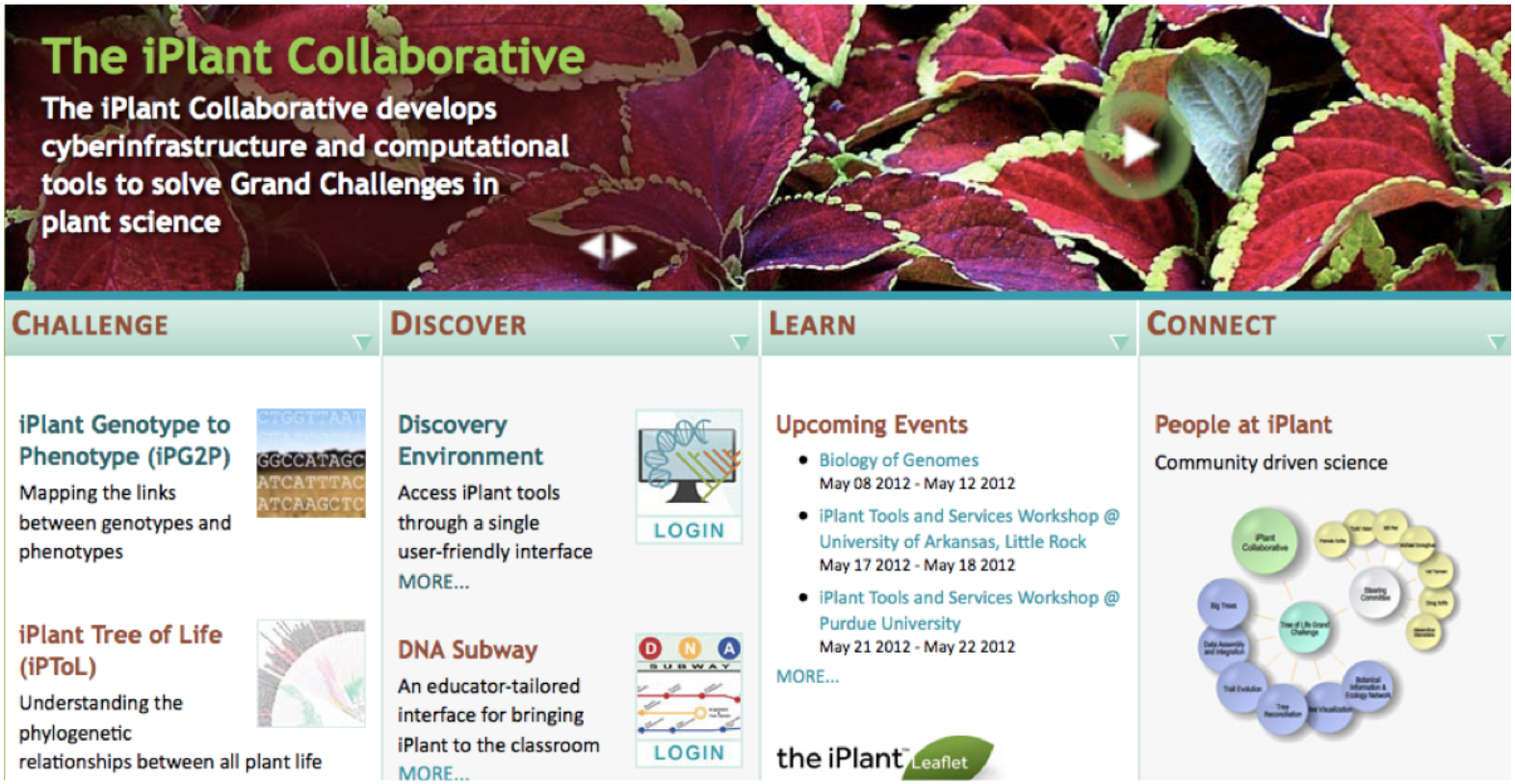
Outil collaboratif



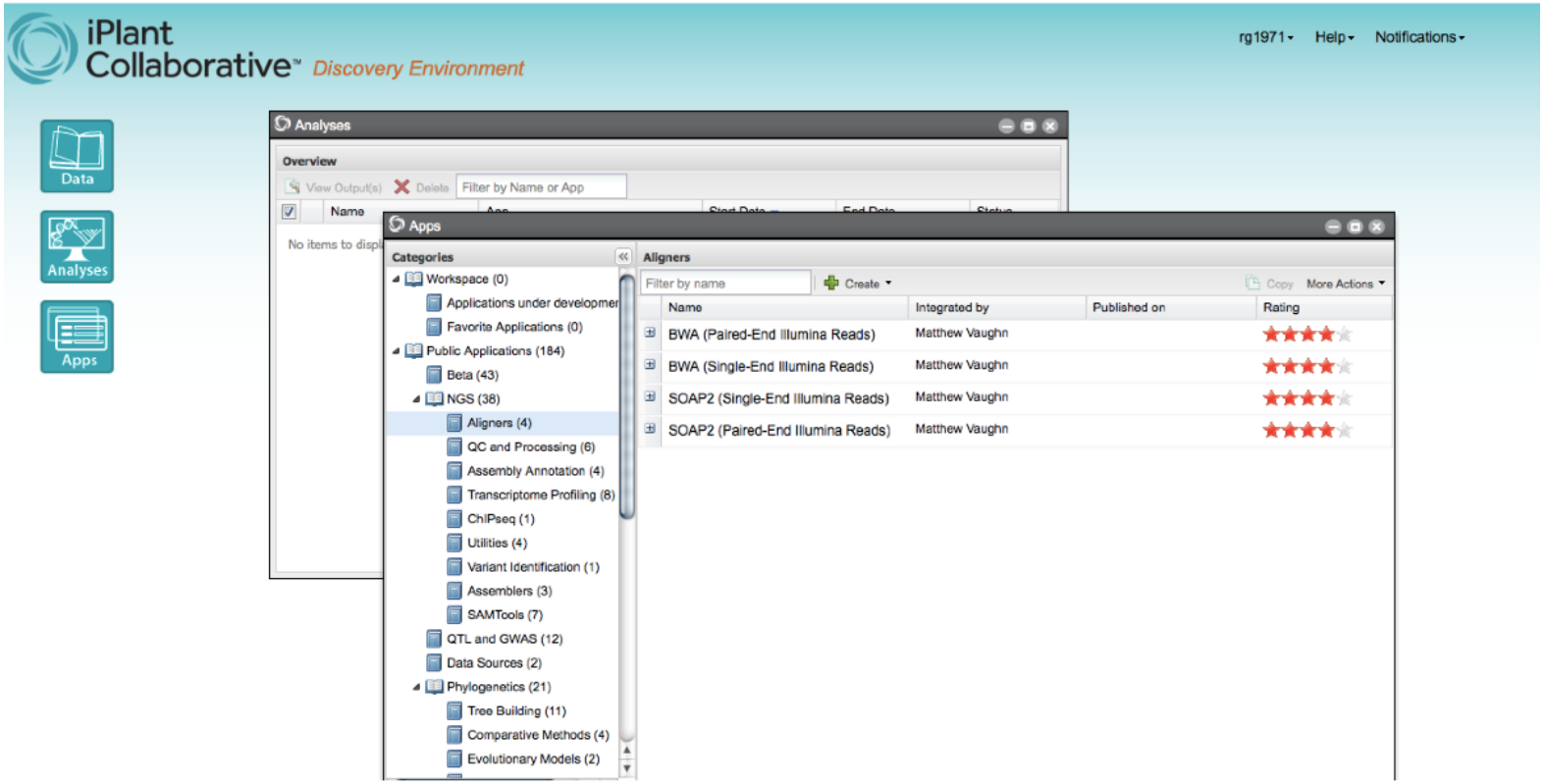
Outil collaboratif



Outil collaboratif



Outil collaboratif



Un site utile pour les analyses NGS