

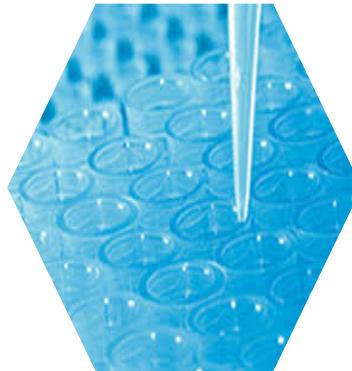
Introduction à la bioinformatique

Christine Tranchant-Dubreuil
Équipe RICE, UMR DIADE

- Approche ***in silico*** de la biologie
- Complémentaire aux approches classiques de la biologie



In vivo



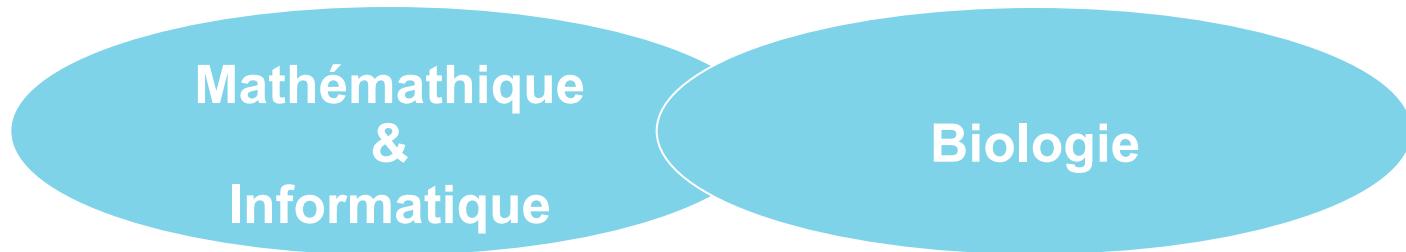
In vitro



In situ

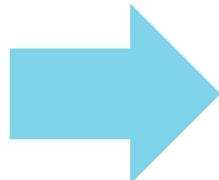
Qu'est ce que la bioinformatique?

- Approche *in silico* de la biologie
- Complémentaire aux approches classiques de la biologie
- **Science interdisciplinaire**



Qu'est ce que la bioinformatique?

- Approche *in silico* de la biologie
- Complémentaire aux approches classiques de la biologie
- **Science interdisciplinaire**



Formaliser des problèmes de biologie moléculaire

Proposer et développer des modèles, méthodes et outils

Qu'est ce que la bioinformatique?

- Approche *in silico* de la biologie
- Complémentaire aux approches classiques de la biologie
- **Science interdisciplinaire**
- 3 principales activités

Acquisition & organisation des données

Conception de logiciels

Analyses des résultats des logiciels

Découverte des lois de l'hérédité

G. Mendel
1866

Identification de la nature chimique de l'ADN

O. Avery, C. McLeod & M. McCarthy
1944

Découverte de la structure en double hélice de l'ADN

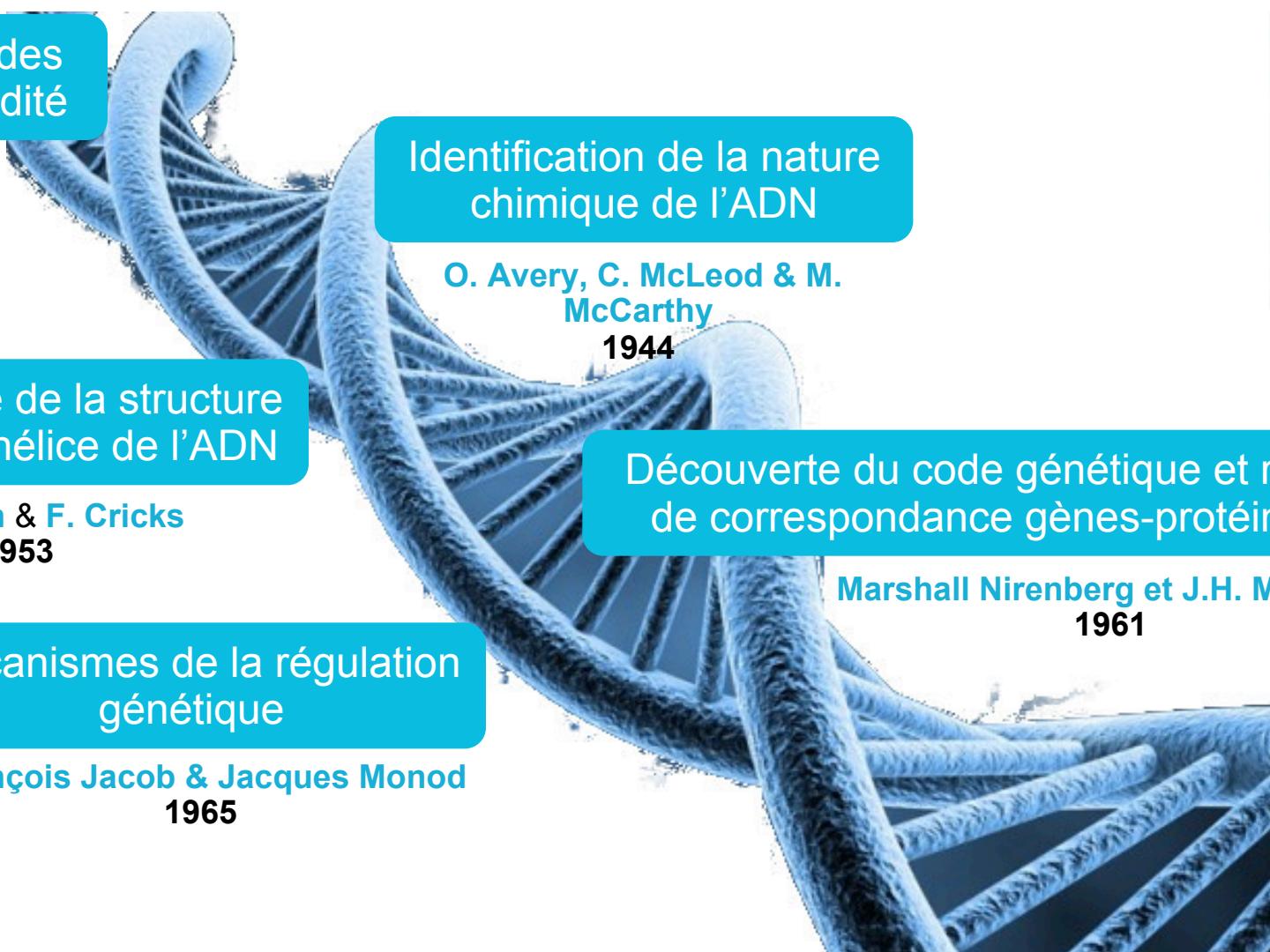
J. Watson & F. Cricks
1953

Découverte du code génétique et règle de correspondance gènes-protéines

Marshall Nirenberg et J.H. Matthaei
1961

Mécanismes de la régulation génétique

François Jacob & Jacques Monod
1965



1970

Programme alignement global de séquence

Algorithme de Needleman & Wunsch

Micro-ordinateurs

1^{ère} suite logicielle (Staden)

1977

Séquençage ADN

F. Sanger / Maxam & Gilbert

1977

EMBL, GenBank

1980

Programme alignement local de séquence

Smith & Waterman

1984

Amplification ADN (PCR)

Karry Mullis

1985

Programme alignement local de séquence FASTA

Pearson & Lipman

1987

1^{er} séquenceur automatisé commercialisé

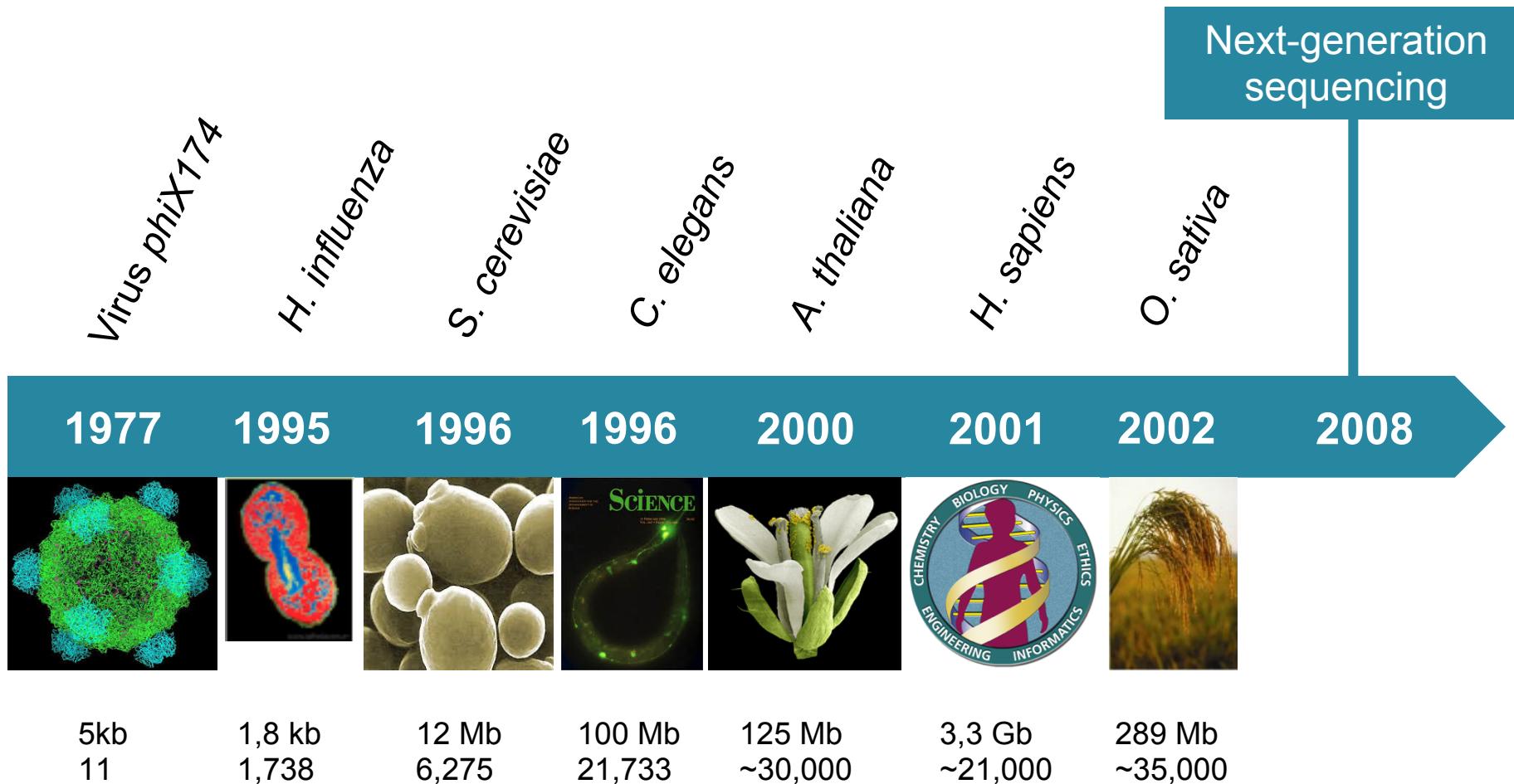
Société Applied Biosystems

1990

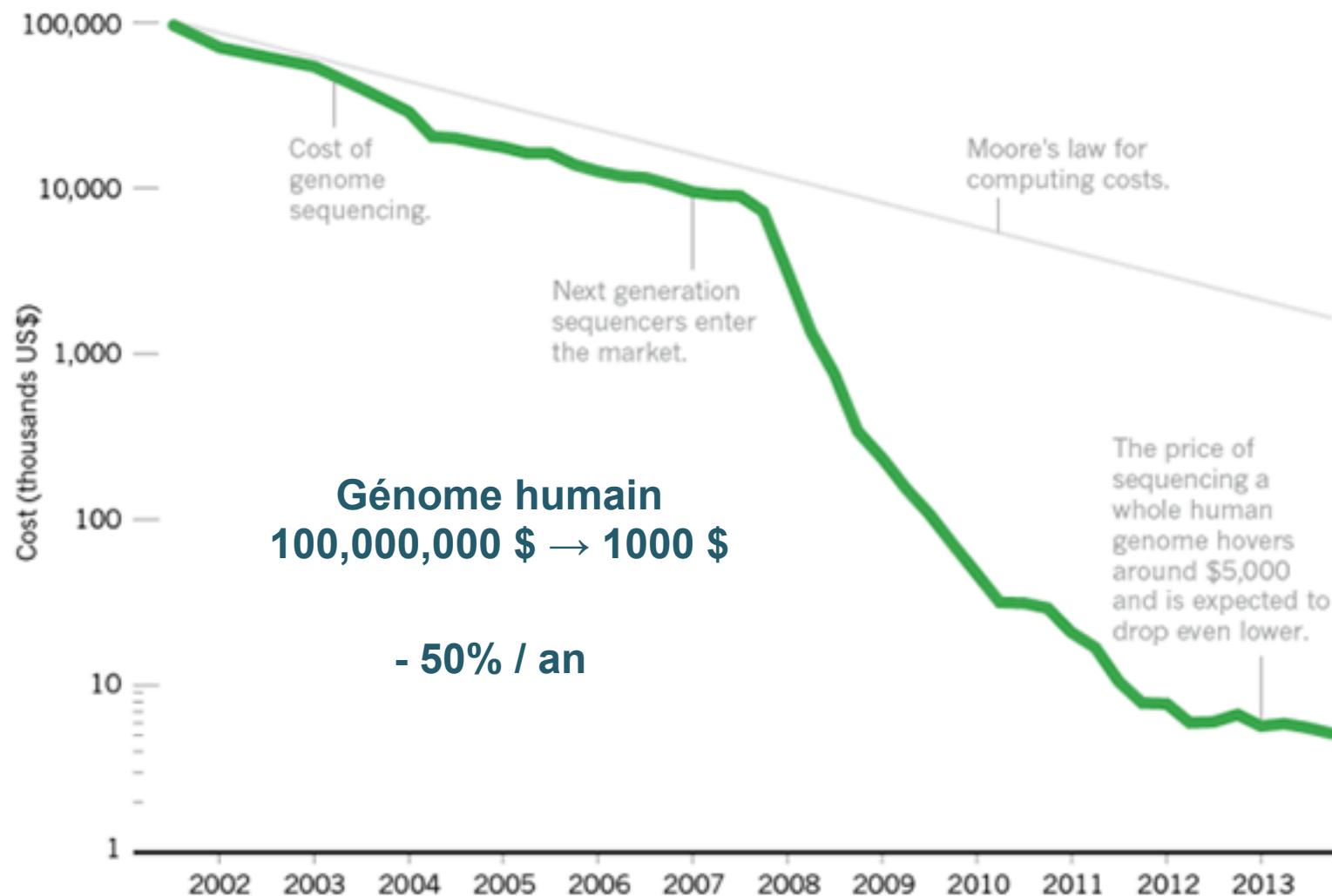
Programme alignement local de séquence BLAST

Altschul et al.

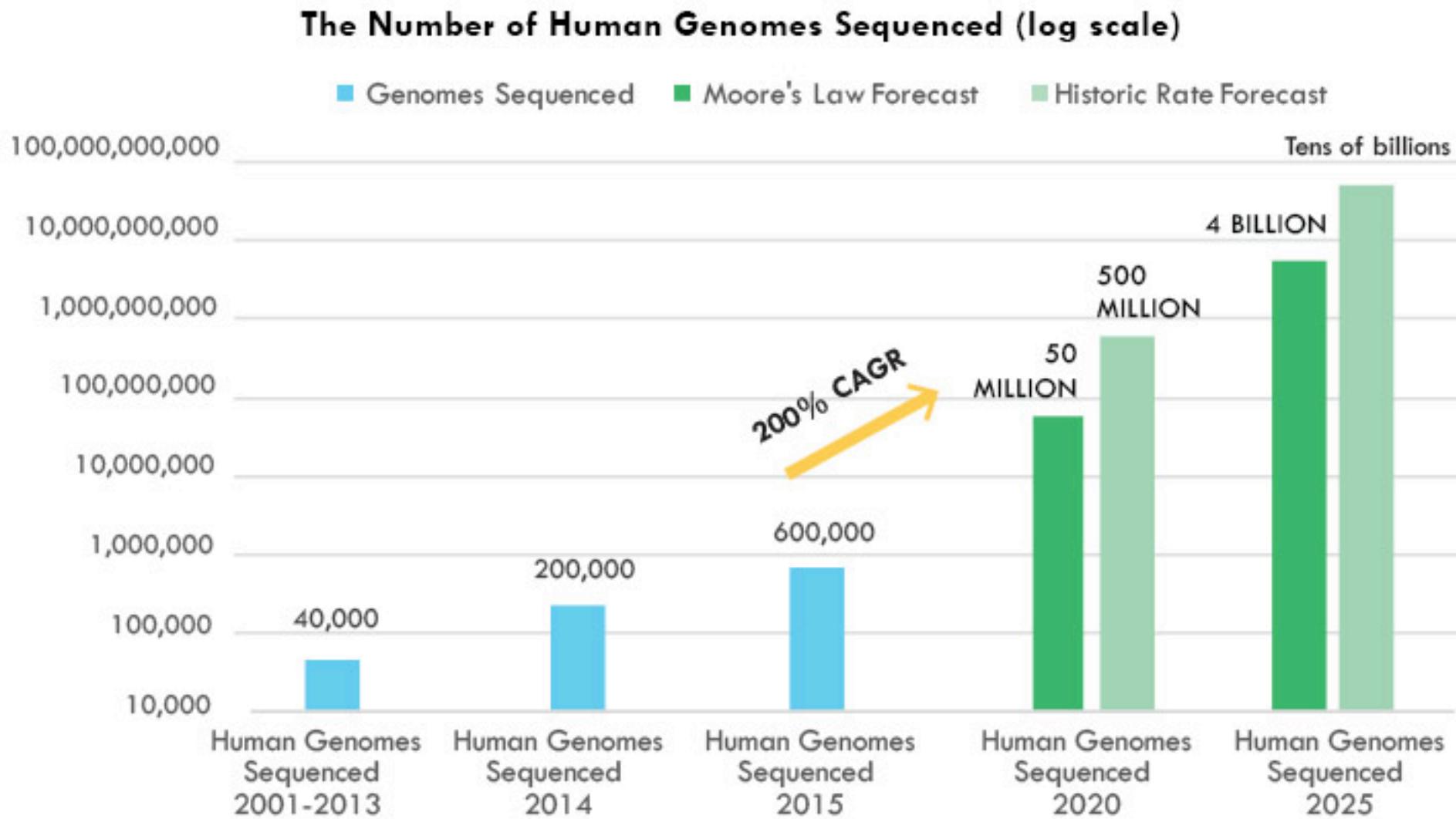
Evolution du séquençage des génomes



Révolution des technologies de séquençage



Révolution des technologies de séquençage



Déluge des données de séquençage

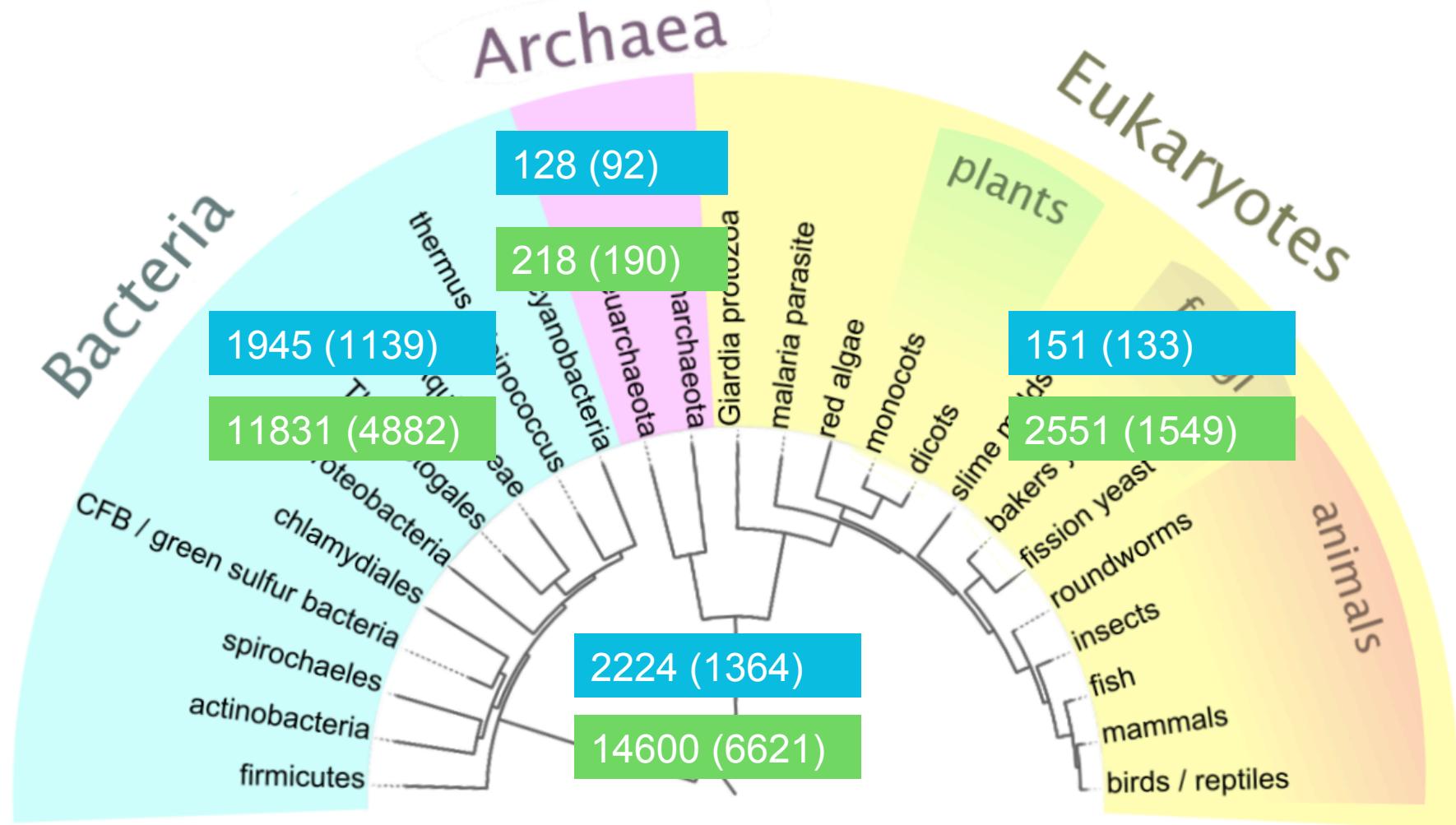
Genome complet

2012 / 2010

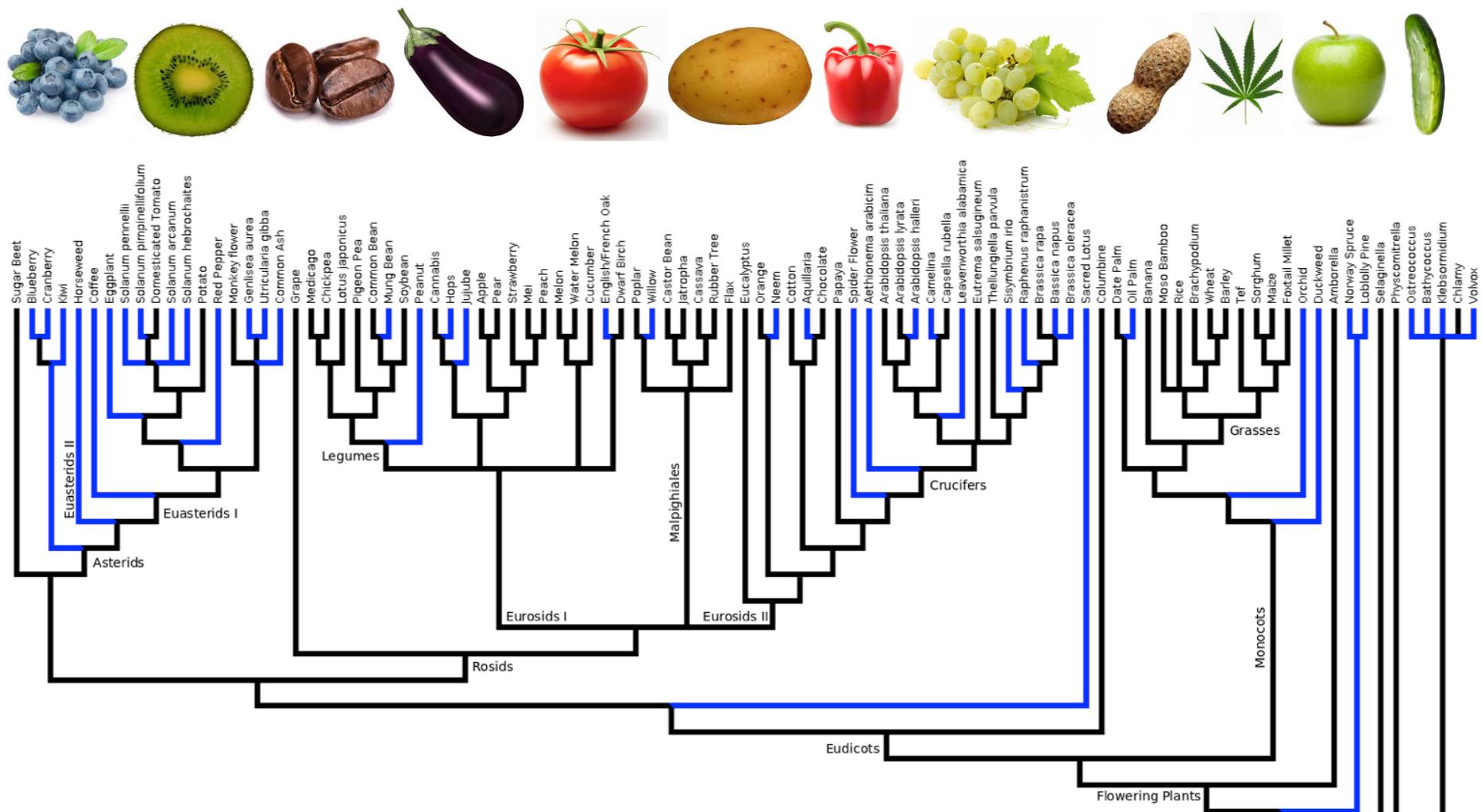
Draft genome

2012 / 2010

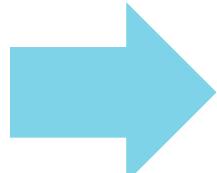
Bilan projets génomes



Déluge des données de séquençage



- Une masse de données génomiques publiques : banques nucléiques, protéiques, génome ...
- De nombreuses ressources web disponibles :
 - Rechercher un gène d'intérêt séquencé dans des espèces proches
 - Définition *in silico* de marqueurs de type SSR, SNP
 - Annotation d'un BAC, scaffold contenant notre gène d'intérêt
 - Génomique comparative



Bioinformatique incontournables pour exploiter ces volumes de données publiques

Banques nucléiques



GenBank



Banques protéiques



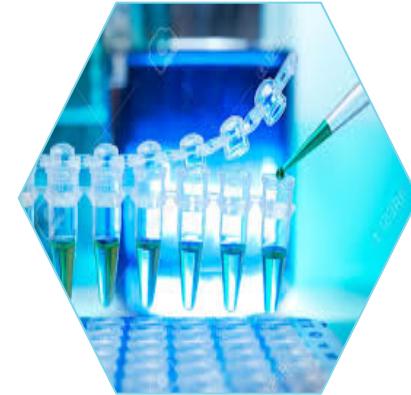
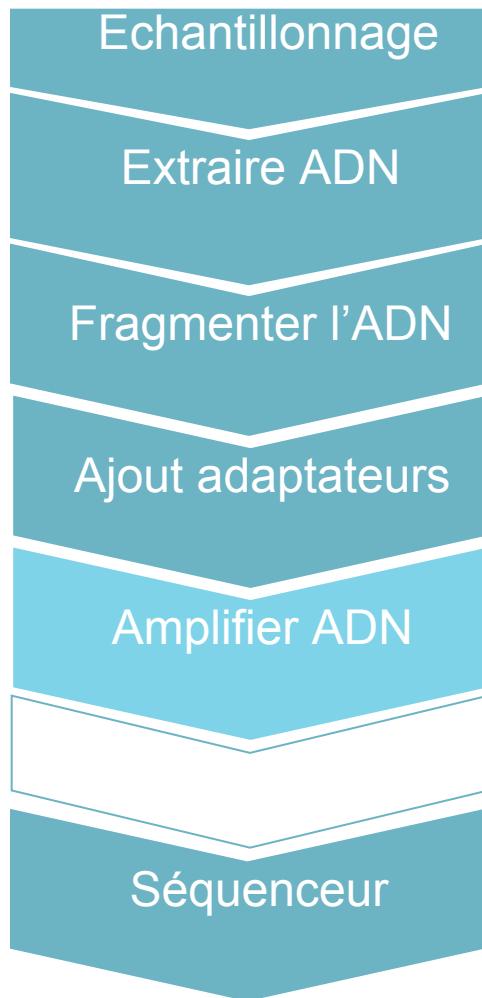
TrEMBL : traduction automatique EMBL

Prosite : familles & domaines protéiques

Autres banques :

Banques de structure, dédiée à un organisme, à un type de séquence

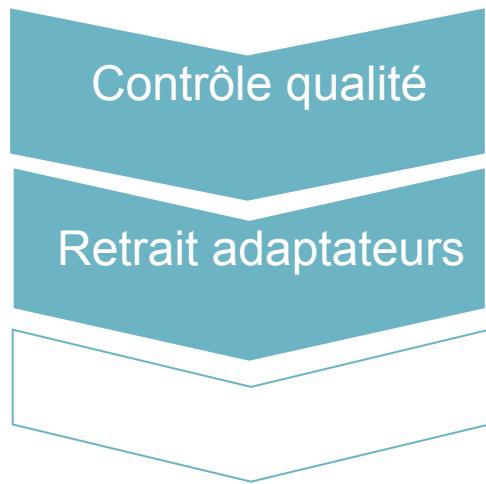
Les différentes étapes du séquençage



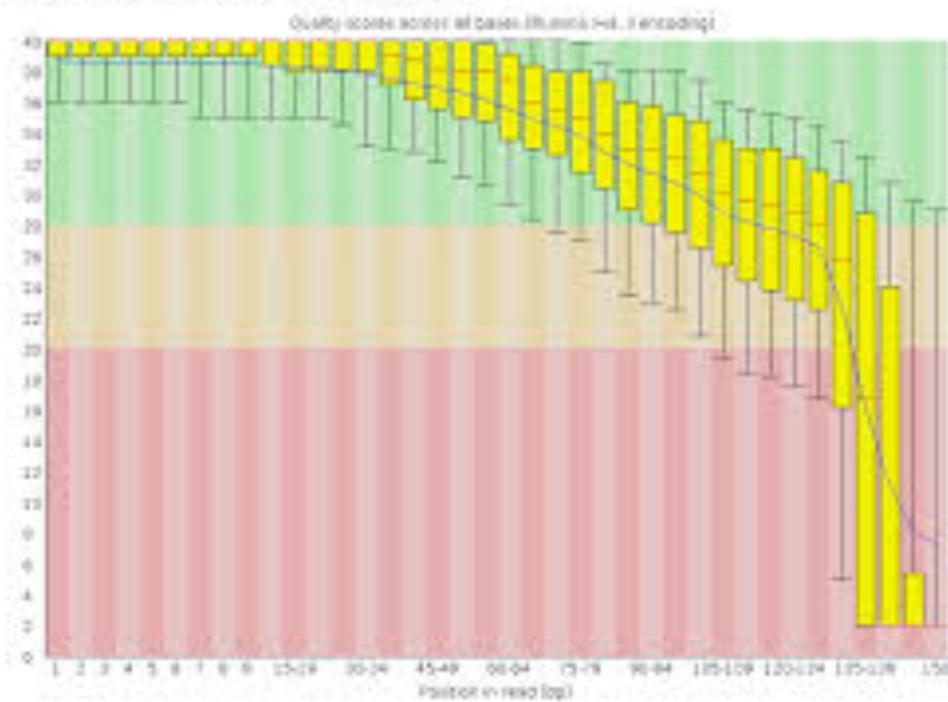
1 milliard pb / réaction



Raw reads

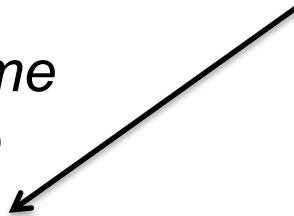


Per base sequence quality



« Reads » séquencés

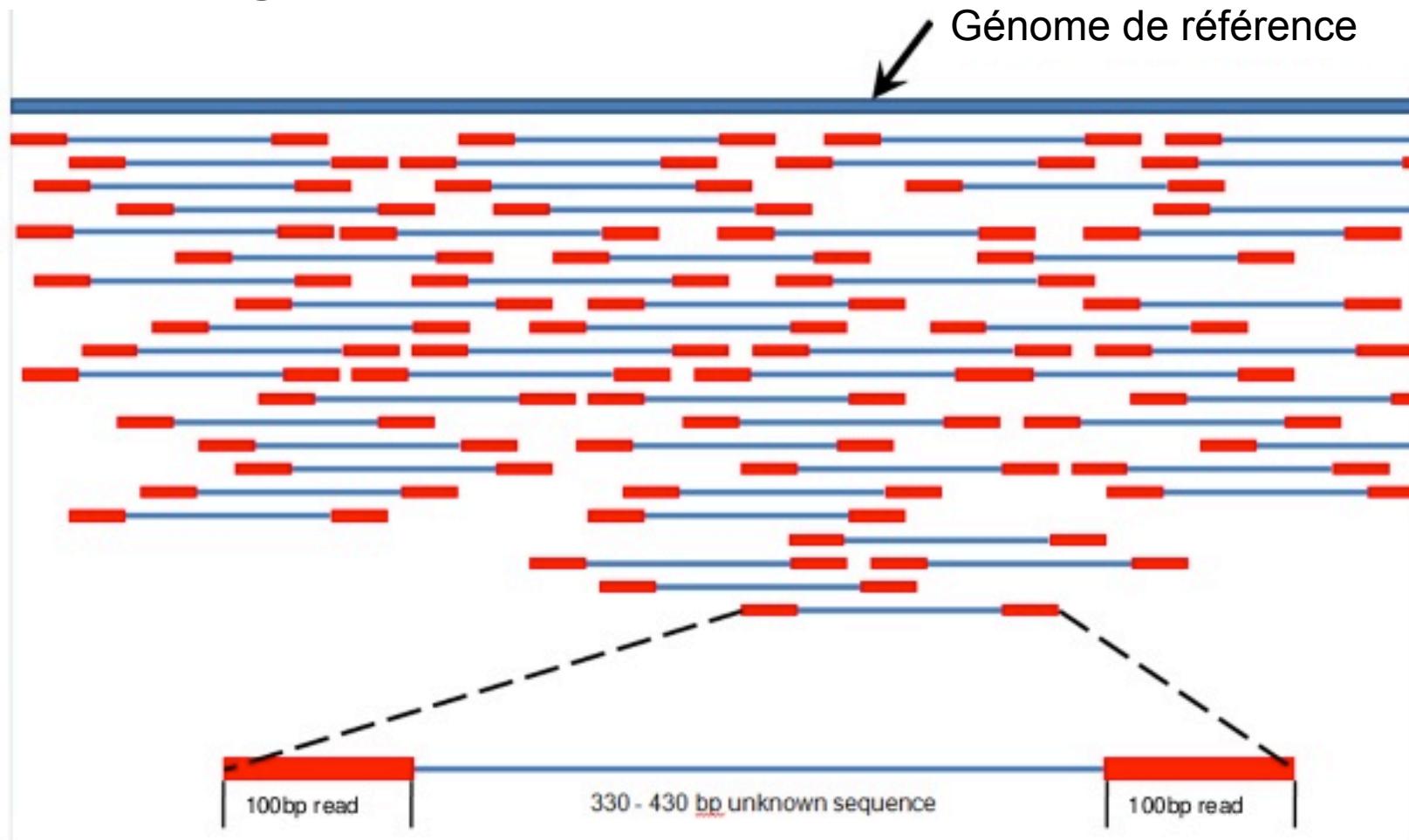
*On a un génome
de référence*



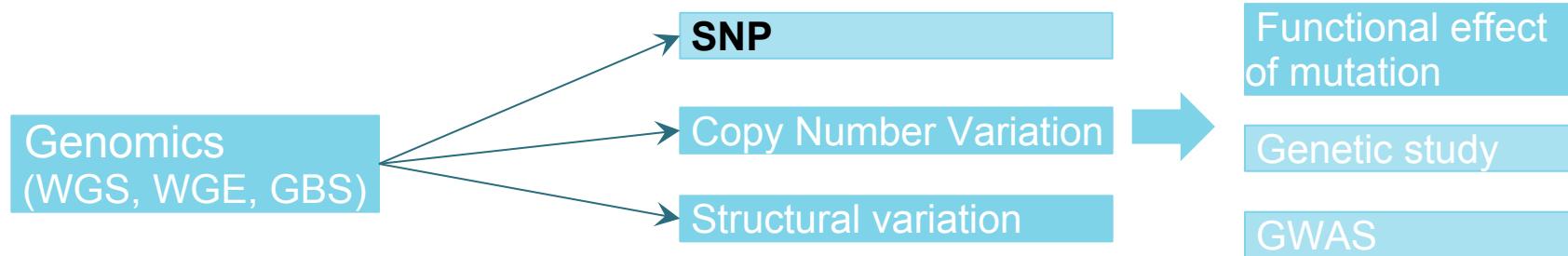
Mapping/alignement

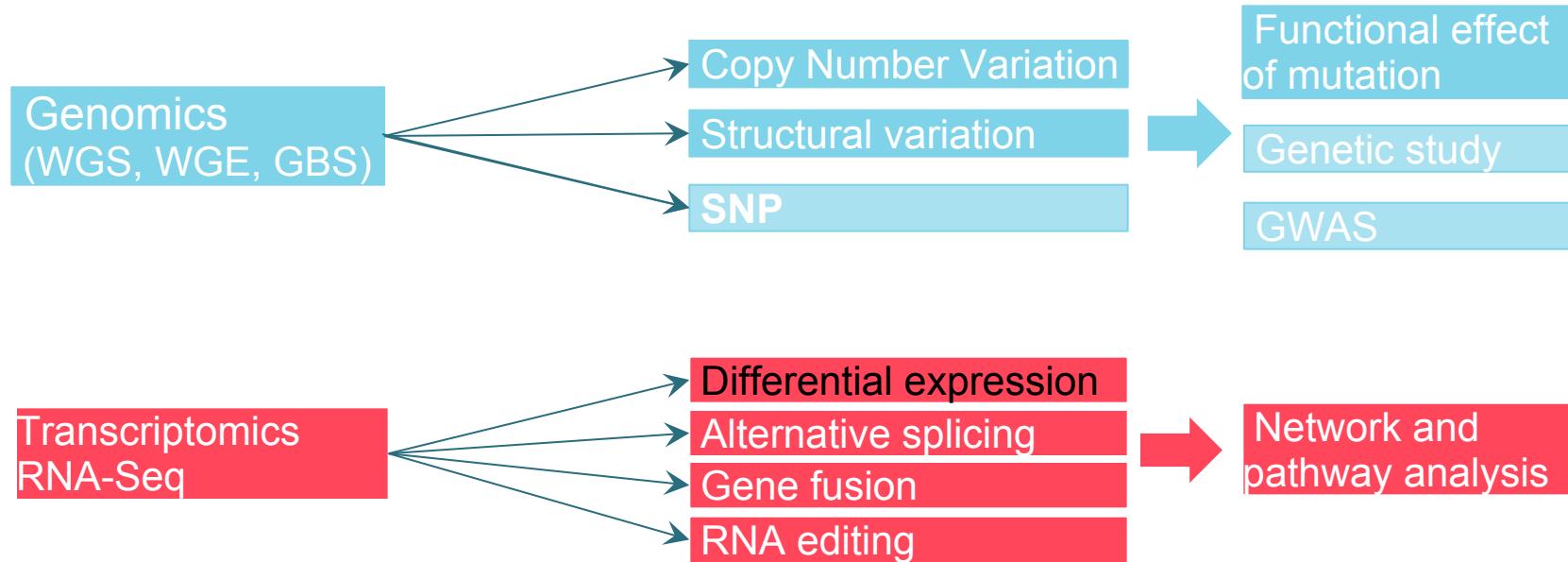
Contre le génome de référence

Si on a un génome disponible

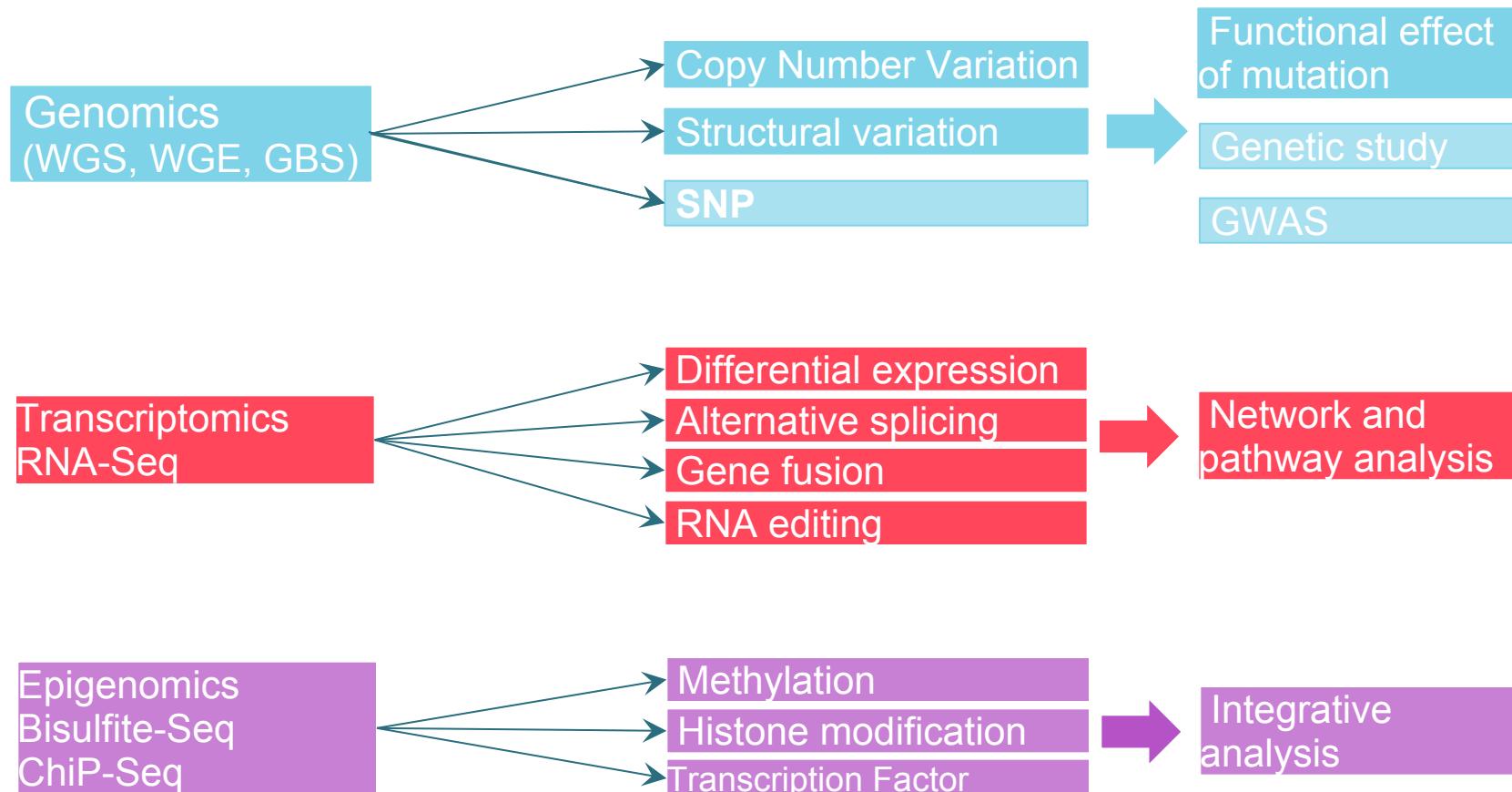


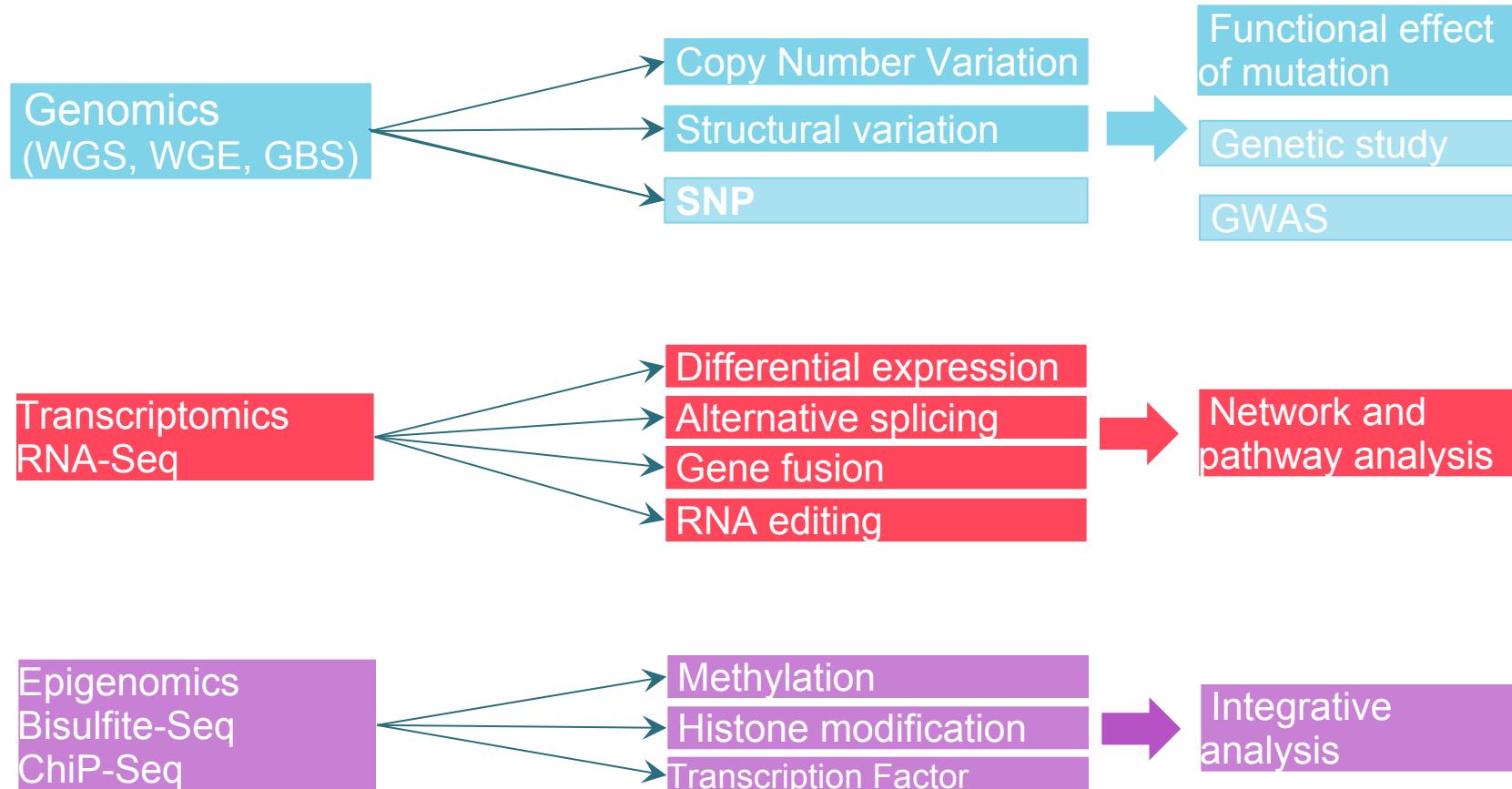
Et après le mapping?





Et après le mapping?





Autre échelle : étude de métagénomique

« Reads » séquencés

*On a un génome
de référence*

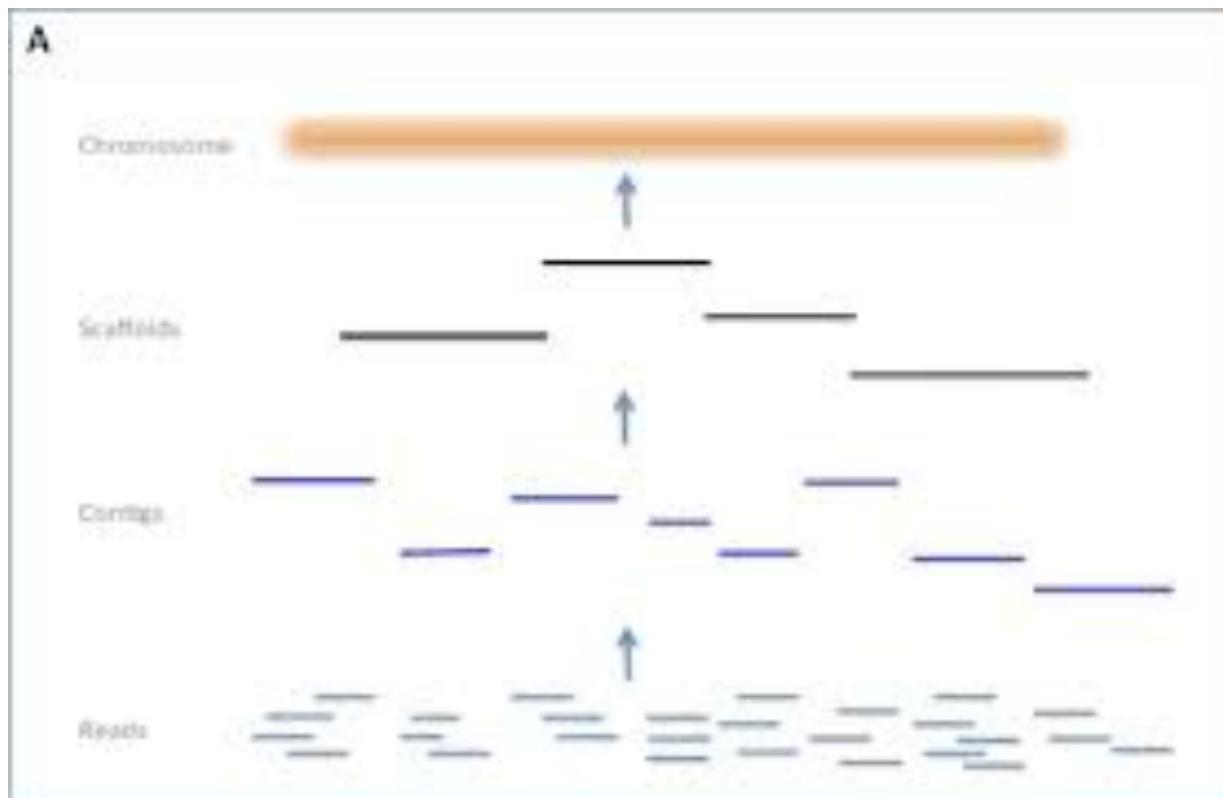


Mapping/alignement
Contre le génome de référence

*On n'a pas un génome
de référence*

Assemblage du génome

Si on n'a pas de génome, étape d'assemblage



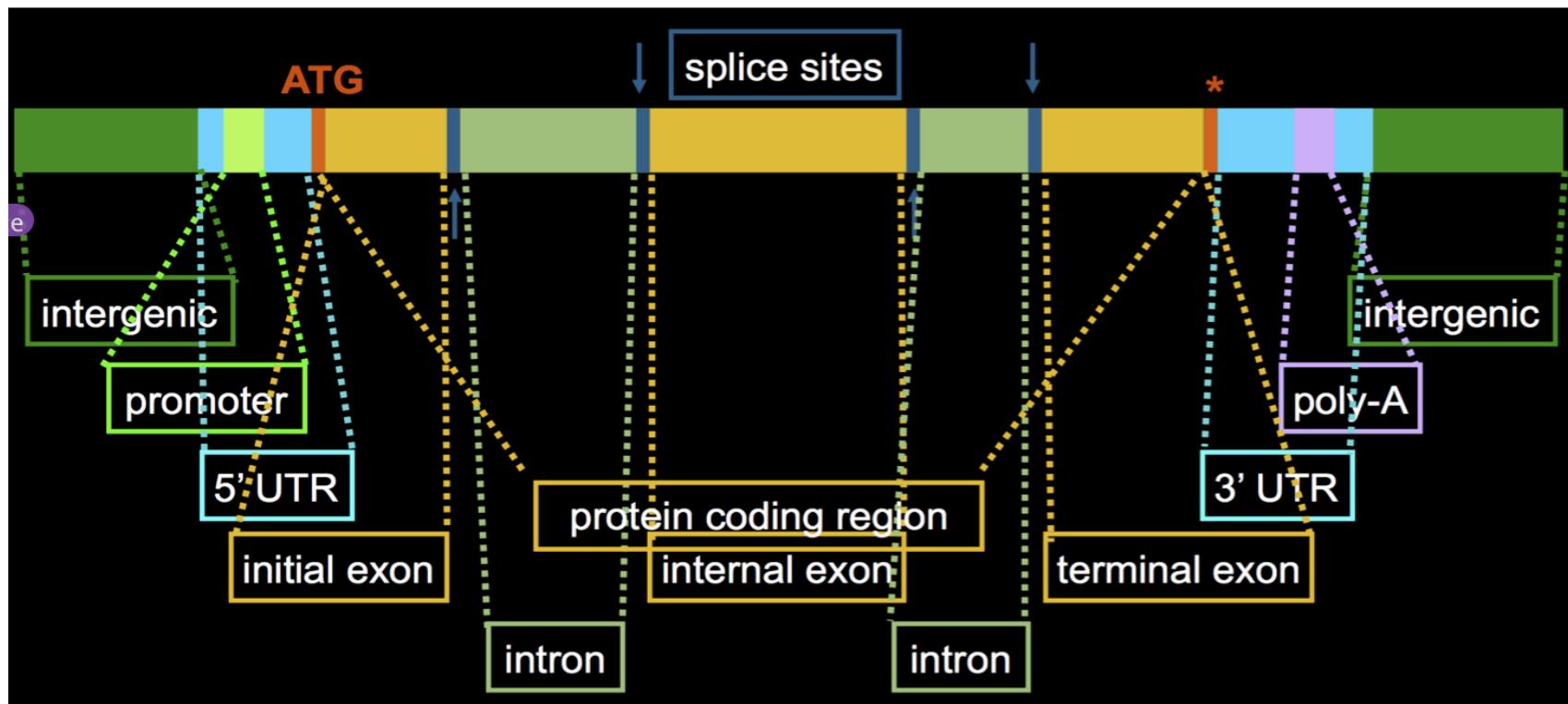
CAACTAGCAGCTAACCGAGCAGAGAAAAATTAAATGAGATTGAAAACAGCTGCAAGCACAAGAAAATTCTGTACAAAGCTCCATCC
AAGAATTGAGAGAAGAACAAAGTCAGCTCACTAGCTCCATCACTGCACCAAATAATAAAATTGATAAGCTAATGGAAGAGGCCAAA
GGAAGATCAATCTGGACCTTCATCTTCAAAACCTCAACCTAGACTGACCTTAGACTAAACCTAACCTAACGCTTAGACTGAAACC
CATGTTTCAAAATATAATCTTGTACTAAGCTGAGTTCATGATTATTGCCACACTTGTATGTACAGATACCAATTGATAACAATTG
CAATTCTCTGATTATCAAATGTACTTATGCAATGAATGATTACCTTCTGAAACATATGTCTCATTGGTTATATACTGCTCTG
ATACATATTGCATCTGATACAAATCAAATCTGTAATGGTATCAAATTCTTAATGTTGGTATCAGATATTATTTAACCTCCTCCT
TTGTTGATGACAAAAAGGGGGAGAAATATAAAAAGAATTCTCATCTGAGAAATGAGAAATATGAAAAAGAATTCTAATCTGA
GAAATAAGAAATGTCAAAAGAATTCAGAATACAAATCTCTGCACTAAAGCTAATATATAAGGGGGAGATTAAATTATGATAAAGAAC
AAGAACTCTGCATAATTGAATGTTGAATTGATAGGGGGAGACCTTCTCCATAATCTGTTGAAATTATGCATATATCTCTGAAC
GTCATATTGCTCTGAACCTCATTATAGAATTCTAATAATTGCTGTGAAATTATTGTAAAATTATTTAActCAATGGTTGTCAT
CATAAAAAAATGGGGAGATTGTCAACCCCTAAAGGATGAATTNN
NNNNNNNNNNNNNNNNNTGCTGAAAACCAACTGGGTTATTGTGATCGCAGAAAACAAAAGGTTTCGATTGGTTGCCT
GAGAAAACCAACTGAATTGATTGTGAGCCCAGAAACATCAGGCTGTAATCTGCTGGGTTAGTGAATCTCAAGCTAGG
CTTGAGGAGTGGACGTAGGTGCTGGAGTGCATCGAACCACTATAATCTGGTGTGATTGTGCTTCTCTCCTTC
CTCTGCATATTCTGACATTCTCATAACTTATTCACTGTCAATTGTACATTCTTATTCCGCTGCTCAATCTTAAAATAAAAGT
AACTCATACTCTTCACGTTTAACTTAACTATTAAAGACCCCCAATTCCCCCCCCCTTGGGTTGCACCTCTGGCA
CAATGAGCAATACCAATATAGTATTGAACGATACATATACACTTATTATAAGCTAAATGAATGTGGAGAGTGATAAACATATT
CCAATACAAAGGAAAATTATAGTTAGCCCATGAGCATTCTGGTTGATGATGATCATTGTTATATCTCCTGGATTCTGG
TACATTCCGTTGCTTACCTTCTATTGGCAAAATCTCCTCCATTGTTCTATAATTGGACTAAAATGAGTTAGATTAAATCTG
ACCGAATCTATTCCATATCTGAAGTGTATCAGATTGAAACACAAATTGATATCCGAATTGATGTGGATCCGAATTGGATTCTGG
AATACTTTCGAATCCAAATCTAGATATCCAGTAAAAAAATTCAAAAATAACTGCATAACCTAGGATTCCAACTAGAGATATA
GTTTGAATGGAATGTCTAACACCAAGATACAATTATTCCTATCTTAAATGAGTTATTGTTATGTATTCA
TCATTAAATATCATATAATGTATAGCATGATATTCTATTAAATCACTATTAAATTTATAAAACTTCTCTTGTATA
ACATACTTTATTATTATAATGTGTTGAAACTTATTGTTAAATTTAAATTTAAATTTAAATTTAAATTTAAATTTAA
TTGCTTACAAGCTTATAAGAATATATAATTCAACATTATTATTGCTATAATTAAATTTAAATTTAAATTTAAATTTAA
CTAGTATTGTTATGTTGATTCTTATCTTAAATGATTGTTATTAAATTTAAATTTAAATTTAAATTTAAATTTAAATTTAA
AAGGTCAAGTCTCATGTCAGTTAATAACCATACTTAGATCCTGCTTACTGCGTATCGTGTGCAATTGATACAAACATG

- Distinguer régions codantes et non codantes
- Réaliser par des programmes informatiques
 - Certains gènes échappent à la détection
 - Certains gènes prédits ne correspondent pas à de vrais gènes
 - Les limites précises du gène sont parfois erronées.

Proportions codant / non codant

Organisme	Génome	Gènes	Codant
Mycoplasma genitalium	0,6 Mb	481	90%
Haemophilus influenzae	1,8 Mb	1.717	86%
Escherichia coli	4,6 Mb	4.289	87%
Saccharomyces cerevisiae	12 Mb	6.286	72%
Caenorhabditis elegans	97 Mb	19.000	27%
Arabidopsis thaliana	120 Mb	27.000	30%
Drosophila melanogaster	165 Mb	16.000	15%
Homo sapiens	3.200 Mb	31.000	3%

Les gènes et leurs produits : Transcrits, protéines



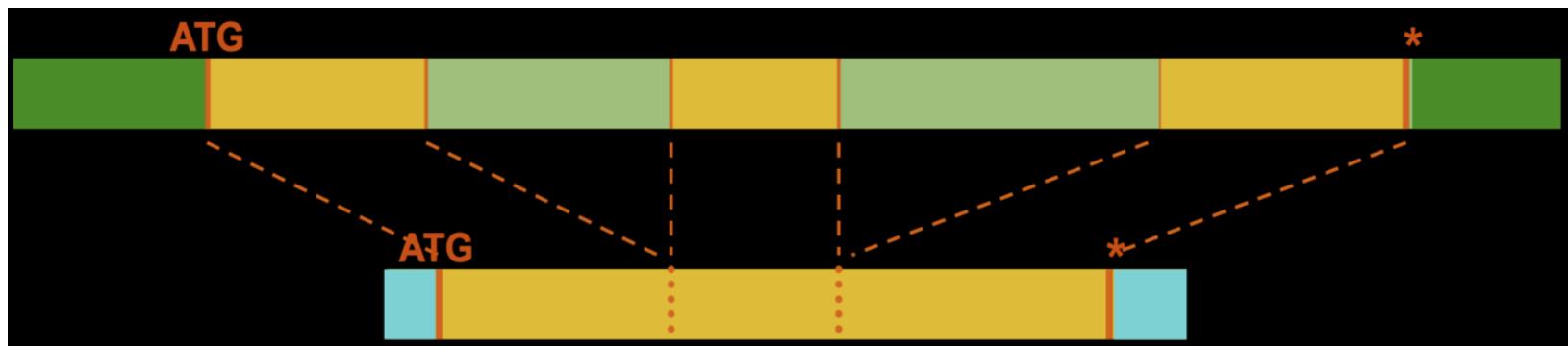
ALIGNEMENT

Méthodes expérimentales

Séquençage de
transcrits issus de
l'organisme
séquencé

Méthodes expérimentales : utilisation de transcrits (cDNA) complets provenant du même organisme

- Transcrits et protéines => évidence directe
- Nécessite des données expérimentales pour chaque gène



ALIGNEMENT

Méthodes expérimentales

Séquençage de
transcrits issus de
l'organisme
séquencé

Méthodes comparatives

Transcrits provenant
du même organisme
(données publiques)

Méthodes comparatives : utilisation de transcrits (cDNA) complets provenant du même organisme

- Comparaison aux séquences d'ESTs disponibles
- Traduction de la séquence génomique en protéines connues (banque de données)
- Comparaison aux séquences génomiques provenant d'espèces proches

ALIGNEMENT

Méthodes expérimentales

Séquençage de transcrits issus de l'organisme séquencé

Méthodes comparatives

Transcrits provenant du même organisme (données publiques)

PREDICTION

ab initio

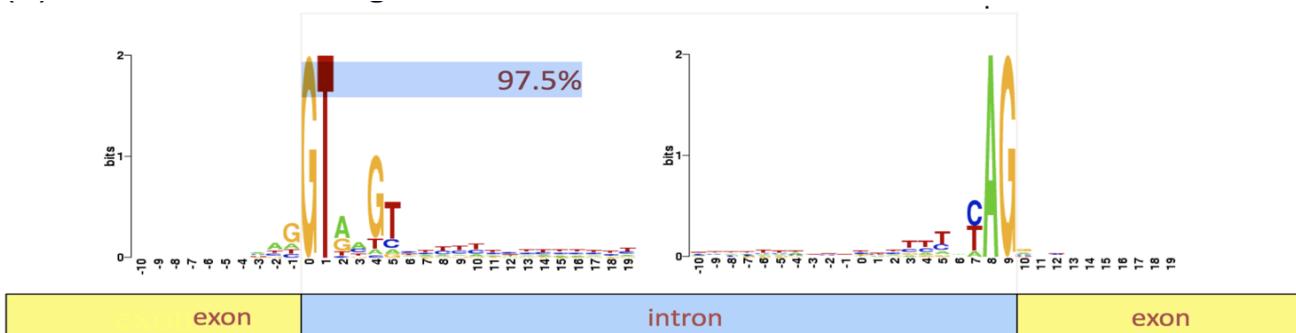
Méthodes ab initio : prédiction de la structure des gènes basées sur des modèles de gènes

1. Recherche des particularités communes à tous les gènes de notre génome
2. Détection sur le génome

Méthodes basées sur les prédictions

Méthodes ab initio : prédiction de la structure des gènes basées sur des modèles de gènes

1. Recherche des particularités communes à tous les gènes de notre génome
2. Détection sur le génome



ALIGNEMENT

Méthodes expérimentales

Séquençage de transcrits issus de l'organisme séquencé

Méthodes comparatives

Transcrits provenant du même organisme (données publiques)

PREDICTION

ab initio

METHODE INTEGRATIVE

Eugène, Maker

Méthodes basées sur les alignements

Comment comparer des séquences ?

Comment comparer des séquences?

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

Méthodes basées sur les alignements

Comment comparer des séquences ?

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

Solution 1 – Algorithme Smith-Waterman / alignement global

Compare votre séquence à chaque séquence de la banque (600 aa vs 85 millions séquences)

+ : Séquence la plus similaire

- : temps de recherche

Progammation dynamique (exacte)

12 millions de séquences, 350 AA/seq

Temps = 0.035 s x 12 millions = 118 heures / 5 jours

Méthodes basées sur les alignements

Comment comparer des séquences ?

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

Solution 2 – Algorithme Smith-Waterman / alignement local - BLAST

Faire une pré-sélection sur les séquences puis les aligner avec SW

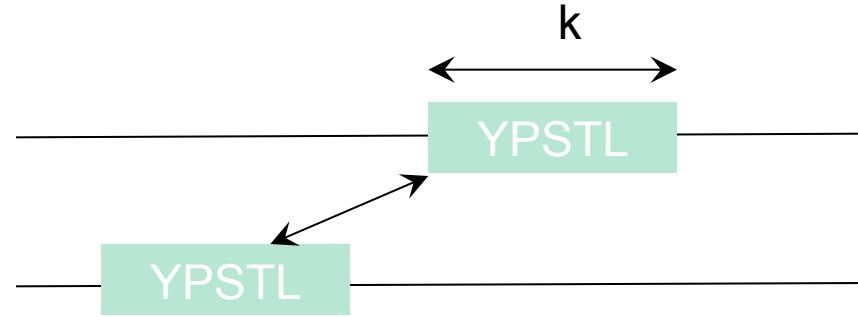
+ : rapide et efficace

- : risque de passer à côté de la perle

Heuristique

Ma séquence

Une séquence
de Genbank

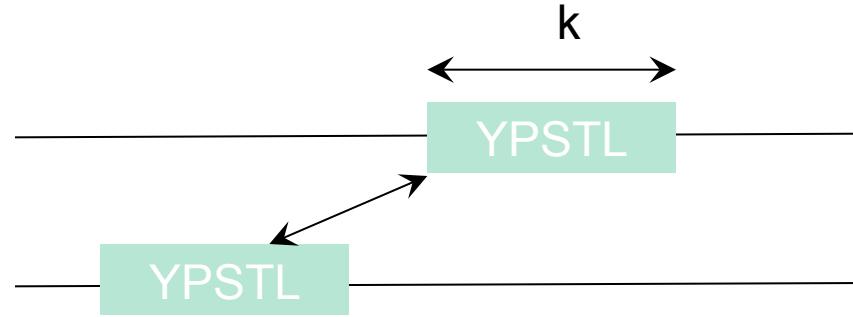


Ne retenir que les séquences partageant au moins un mot de longueur k avec ma séquence

Pourquoi cette sélection est elle si rapide ?

Ma séquence

Une séquence
de Genbank



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

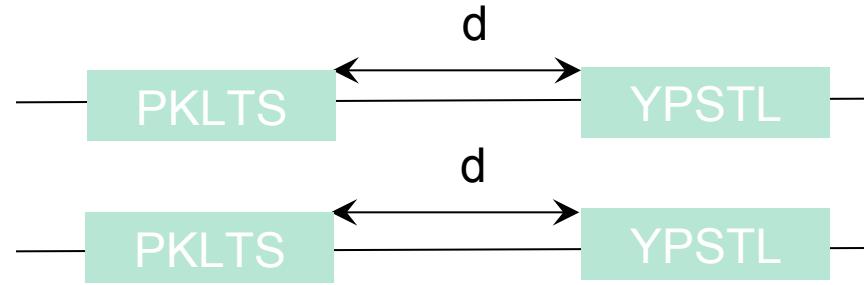
Pourquoi cette sélection est elle si rapide ?

YYYY
Accession1
Accession2
YPSTL
Accession1
Accession2

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

Ma séquence

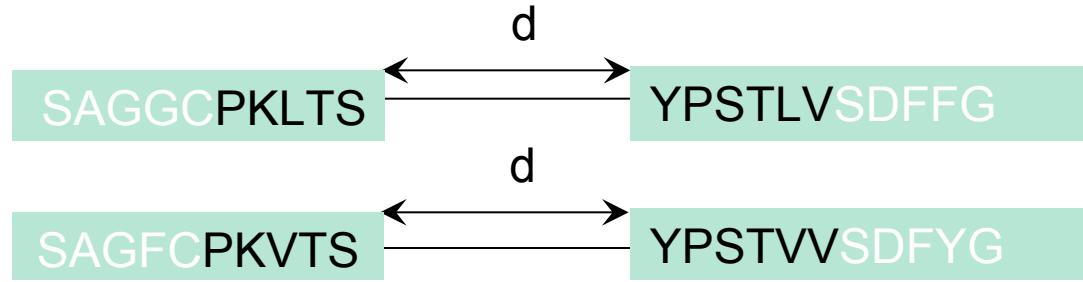
Une séquence
de Genbank



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

Ma séquence

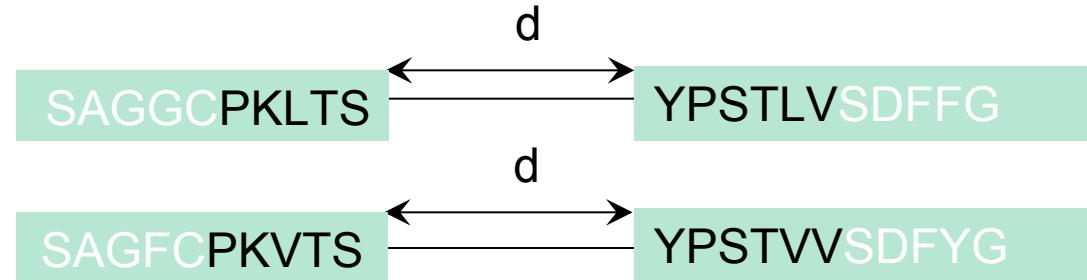
Une séquence
de Genbank



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

Ma séquence

Une séquence
de Genbank



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

S A G G C P K L T S
| | | | | | | | | |
S A G F C P K V T S

BLAST® » blastx

Translated BLAST: blastx

[blastn](#) [blastp](#) **blastx** [tblastn](#) [tblastx](#)

Enter Query Sequence

BLASTX search protein databases using a translated nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) 

[Clear](#)

Query subrange 

1. Requête : votre séquence

Or, upload file

Choisissez un fichier Aucun fichier choisi 

Genetic code

Standard (1) 

Job Title

Enter a descriptive title for your BLAST search 

Align two or more sequences 

Choose Search Set

Database

Non-redundant protein sequences (nr) 

Organism
Optional

green plants (taxid:33090)

Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 

Exclude
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query
Optional

 [Create custom database](#)

Enter an Entrez query to limit search 

2. Banque de séquences

BLAST

Search **database Non-redundant protein sequences (nr)** using Bla

Show results in a new window

3. Top, c'est parti!

de query)

dbj|AK287457.1| (1995 letters)

RID R0DUSNGN01R (Expires on 06-28 04:13 am)

Query ID gi|156764072|dbj|AK287457.1|

Description Oryza sativa Japonica Group cDNA, clone: J043019L17, full insert sequence

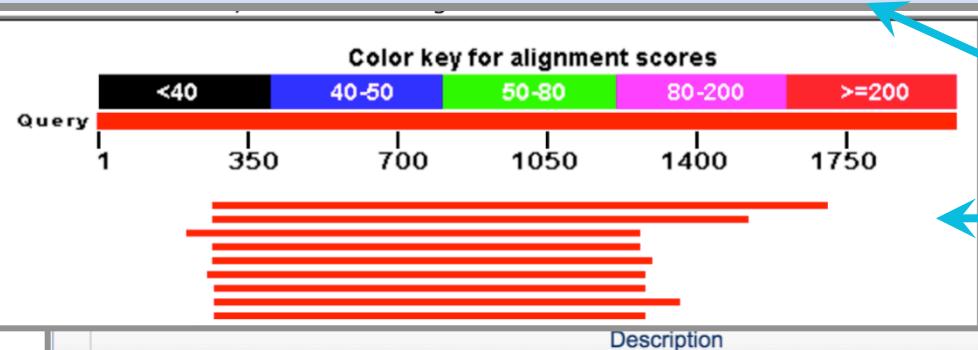
Molecule type nucleic acid

Query Length 1995

Database Name swissprot

Description Non-redundant UniProtKB/SwissProt sequences

Program BLASTX 2.4.0+ [Citation](#)



1. Récapitulatif de la requête

2. Graphique des résultats

		Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN-1; AltName: Full=Kinesin-1-like protein PSS1; AltName: Full=Pollen semi-sterility protein	835	835	71%	0.0	100%	F9W301.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN-1; AltName: Full=AtKIN-1; AltName: Full=AtPSS1; AltName: Full=Pollen semi-sterility pr	493	493	62%	1e-169	61%	Q8GW44.1
<input type="checkbox"/>	RecName: Full=Phragmoplast orienting kinesin 2; AltName: Full=Kinesin POK2	243	243	52%	3e-68	39%	Q27IK6.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN12A; AltName: Full=Phragmoplast-associated kinesin-related protein 1; Short=AtPAKRP-	238	238	49%	1e-66	39%	Q9LDN0.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN12B; AltName: Full=Phrac						
<input type="checkbox"/>	RecName: Full=Kinesin-like protein FLA10; AltName: Full=Protein						

RecName: Full=Protein HEADING DATE 3A; AltName: Full=FT-like protein A
Sequence ID: sp|Q93WI9.1|HD3A_ORYSJ Length: 179 Number of Matches: 1

Range 1: 6 to 169 GenPept Graphics				▼ Next Match	▲ Previous Match	
Score	Expect	Method	Identities	Positives	Gaps	Frame
298 bits(763)	6e-103	Compositional matrix adjust.	135/164(82%)	152/164(92%)	0/164(0%)	+1

Query 100	RDRDPLVVGRVIGDVLDPFIRSISILRVNNYNSREVNNNGCELKPSQVVSQPRVDIGGDDLRT	279
Sbjct 6	RDRDPLVVGRV+GDVLD F+RS +L+V Y S+ V+NGCELKPS V QPRV++GG+D+RT	65
Query 280	FYTLVMVDPDAPSPSDPNLREYLHLWLTDIPATTGASFGQEIVCYENPRPTVGIHRFVFV	459
Sbjct 66	FYTLVMVDPDAPSPSDPNLREYLHLWLTDIPATTGASFGQEIVCYENPRPTVGIHRFVFV	125
Query 460	LFRQLGRQTYYAPGWRQNFTNDFAEYLNLGLPVASVYFNQCQE	591
Sbjct 126	LFQQLGRQTYYAPGWRQNFTNDFAEYLNLGLPVASVYFNQCQE	169

3. Résumé des résultats

4. Alignements

1. Récapitulatif de la requête

dbj|AK287457.1| (1995 letters)

RID [R0DUSNGN01R](#) (Expires on 06-28 04:13 am)

Query ID [gi|156764072|dbj|AK287457.1|](#)

Description Oryza sativa Japonica Group cDNA, clone: J043019L17, full insert sequence

Molecule type nucleic acid

Query Length 1995

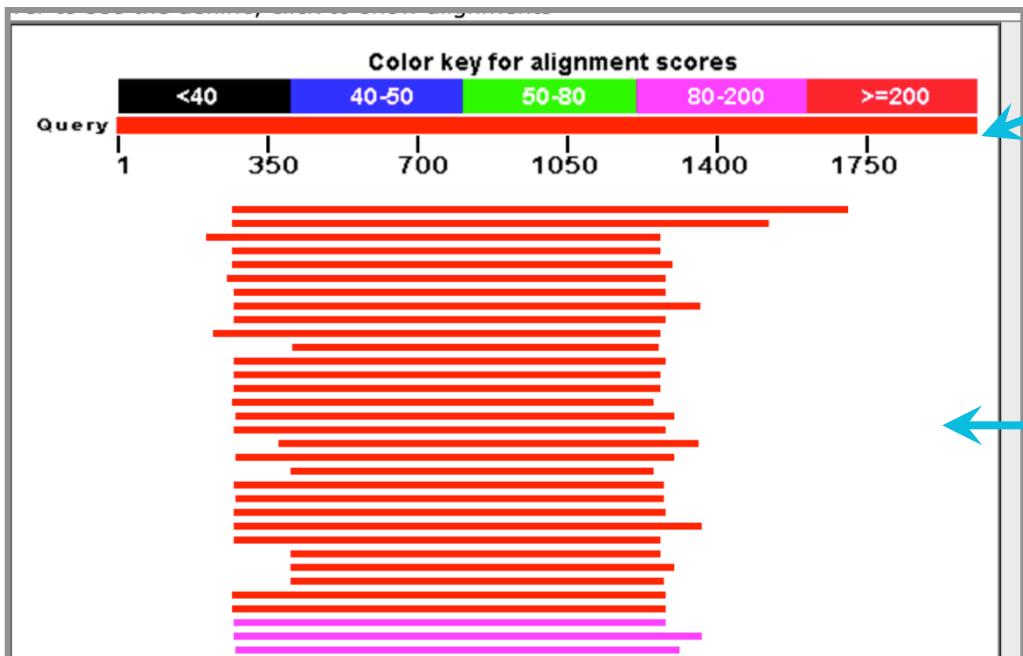
Database Name swissprot
Description Non-redundant UniProtKB/SwissProt sequences
Program BLASTX 2.4.0+ [Citation](#)

Séquence

Banque interrogée

Programme utilisé

2. Graphique des résultats



Séquence soumise

HSP (High Scoring Pair)

1 trait = 1 alignement

Couleur -> score
Longueur -> taille alignement

3. Résumé des résultats

Descriptions

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

All Alignments Download GenPept Graphics

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN-1; AltName: Full=Kinesin-1-like protein PSS1; AltName: Full=Pollen semi-sterility protein	835	835	71%	0.0	100%	F9W301.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN-1; AltName: Full=AtKIN-1; AltName: Full=AtPSS1; AltName: Full=Pollen semi-sterility pr	493	493	62%	1e-169	61%	Q8GW44.1
<input type="checkbox"/>	RecName: Full=Phragmoplast orienting kinesin 2; AltName: Full=Kinesin POK2	243	243	52%	3e-68	39%	Q27IK6.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN12A; AltName: Full=Phragmoplast-associated kinesin-related protein 1; Short=AtPAKRP	238	238	49%	1e-66	39%	Q9LDN0.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN12B; AltName: Full=Phragmoplast-associated kinesin-related protein 1-like protein; Short	233	233	51%	6e-65	38%	Q8L7Y8.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein FLA10; AltName: Full=Protein KHP1	229	229	50%	6e-65	38%	P46869.1
<input type="checkbox"/>	RecName: Full=Phragmoplast orienting kinesin-1; AltName: Full=Kinesin POK1	229	229	50%	1e-63	39%	Q27IK7.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN-7C, mitochondrial; AltName: Full=Mitochondria-targeted kinesin-related protein 1; Flags:	214	214	54%	3e-59	36%	Q8W5R6.1
<input type="checkbox"/>	RecName: Full=Kinesin-like protein KIN-5C; AltName: Full=125 kDa kinesin-related protein	215	215	50%	3e-59	38%	O23826.1

4. Alignments

RecName: Full=Kinesin-like protein KIN-1; AltName: Full=Kinesin-1-like protein PSS1; AltName: Full=Pollen semi-sterility protein 1
Sequence ID: [sp|F9W301.1|KN1_ORYSJ](#) Length: 477 Number of Matches: 1

Range 1: 1 to 477 [GenPept](#) [Graphics](#)

			▼ Next Match	▲ Previous Match		
Score	Expect	Method	Identities	Positives	Gaps	Frame
835 bits(2156)	0.0	Compositional matrix adjust.	477/477(100%)	477/477(100%)	0/477(0%)	+1
Query 2			MSNVTCVRFRPLSHKERKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	RVFYEDAEQS	451	
Sbjct 1			MSNVTCVRFRPLSHKERKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	RVFYEDAEQS	60	
Query 451			DVYNFLAVPIVADAISGTINGIITYGQTGAGKTYSMEGPSILHCNKQKTGLVQRVVDEL	F	631	
Sbjct 61			DVYNFLAVPIVADAISGTINGIITYGQTGAGKTYSMEGPSILHCNKQKTGLVQRVVDEL	F	121	
Query 631			QSLQSSESMAMWSVKLSMVEIYLEKVRDLILSKDNLQIKESKTQGIYISGATEVSIQNS	QSLQSSESMAMWSVKLSMVEIYLEKVRDLILSKDNLQIKESKTQGIYISGATEVSIQNS	811	
Sbjct 121			QSLQSSESMAMWSVKLSMVEIYLEKVRDLILSKDNLQIKESKTQGIYISGATEVSIQNS	QSLQSSESMAMWSVKLSMVEIYLEKVRDLILSKDNLQIKESKTQGIYISGATEVSIQNS	181	
Query 811			SDALECLSEGIANRAVGETQMNLASSRSHCLYIIFSVQQGSTSDERVRGGKIIILVDLAGSE	SDALECLSEGIANRAVGETQMNLASSRSHCLYIIFSVQQGSTSDERVRGGKIIILVDLAGSE	991	
Sbjct 181			SDALECLSEGIANRAVGETQMNLASSRSHCLYIIFSVQQGSTSDERVRGGKIIILVDLAGSE	SDALECLSEGIANRAVGETQMNLASSRSHCLYIIFSVQQGSTSDERVRGGKIIILVDLAGSE	241	
Query 991			KVEKTGAEGRVLDEAKTINKSLSVLGNVNVNALTGKPNHVPRDSKLTRILQDALGGNSR	KVEKTGAEGRVLDEAKTINKSLSVLGNVNVNALTGKPNHVPRDSKLTRILQDALGGNSR	1171	
Sbjct 241			KVEKTGAEGRVLDEAKTINKSLSVLGNVNVNALTGKPNHVPRDSKLTRILQDALGGNSR	KVEKTGAEGRVLDEAKTINKSLSVLGNVNVNALTGKPNHVPRDSKLTRILQDALGGNSR	301	
Query 1171			AALLCCSPSASNAPESLSTVRFGTRTKLIKTPKSISPEVDSIKKPIPDSHGQNQLRDR	AALLCCSPSASNAPESLSTVRFGTRTKLIKTPKSISPEVDSIKKPIPDSHGQNQLRDR	1350	
Sbjct 301			AALLCCSPSASNAPESLSTVRFGTRTKLIKTPKSISPEVDSIKKPIPDSHGQNQLRDR	AALLCCSPSASNAPESLSTVRFGTRTKLIKTPKSISPEVDSIKKPIPDSHGQNQLRDR	360	
Query 1351			IILNKLRLSLKEEDVLLLEELFVQEGLIFDPNYSVADIDSACQDAASQEVSLLTQAveelk	IILNKLRLSLKEEDVLLLEELFVQEGLIFDPNYSVADIDSACQDAASQEVSLLTQAveelk	1530	
Sbjct 361			IILNKLRLSLKEEDVLLLEELFVQEGLIFDPNYSVADIDSACQDAASQEVSLLTQAveelk	IILNKLRLSLKEEDVLLLEELFVQEGLIFDPNYSVADIDSACQDAASQEVSLLTQAveelk	420	
Query 1531			etveeltdeNERlrlgelelaqeaaaaaaaaaradgallgFVPVAISSLRPFGFV	etveeltdeNERlrlgelelaqeaaaaaaaaaradgallgFVPVAISSLRPFGFV	421	
Sbjct 421			etveeltdeNERlrlgelelaqeaaaaaaaaaradgallgFVPVAISSLRPFGFV	etveeltdeNERlrlgelelaqeaaaaaaaaaradgallgFVPVAISSLRPFGFV	1701	
					477	

Query
S é q u e n c e
soumise

Related Information

S u b j e c t
s é q u e n c e
Banque

 Download [▼ GenPept](#) [Graphics](#)

[▼ Next](#) [▲ Previous](#) [▲ Descriptions](#)

RecName: Full=Kinesin-like protein KIN-1; AltName: Full=AtKIN-1; AltName: Full=AtPSS1; AltName: Full=Pollen semi-sterility protein 1
Sequence ID: [sp|Q8GW44.1|KN1_ARATH](#) Length: 465 Number of Matches: 1

Range 1: 1 to 417 [GenPept](#) [Graphics](#)

			▼ Next Match	▲ Previous Match		
Score	Expect	Method	Identities	Positives	Gaps	Frame
493 bits(1269)	1e-169	Compositional matrix adjust.	257/420(61%)	330/420(78%)	8/420(1%)	+1
Query 271			MSNVTCVRFRPLSHKE-RKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	RVFYEDAEQ	447	
Sbjct 1			MSNVTCVRFRPLSHKE-RKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	RVFYEDAEQ	60	

Related Information

Gene - associated gene details

4. Alignements

RecName: Full=Kinesin-like protein KIN-1; AltName: Full=Kinesin-1-like protein PSS1; AltName: Full=Pollen semi-sterility protein 1
Sequence ID: [sp|F9W301.1|KN1_ORYSJ](#) Length: 477 Number of Matches: 1

Range 1: 1 to 477 [GenPept](#) [Graphics](#)

			Identities	Positives	Gaps	Frame
Score	Expect	Method	477/477(100%)	477/477(100%)	0/477(0%)	+1
835 bits(2156)	0.0	Compositional matrix adjust.				
Query	2	MSNVTCVRFRPLSHKERKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	451			
Sbjct	1	MSNVTCVRFRPLSHKERKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	60			
Query	451	DVYNFLAVPIVADAISGTINGIITYQGTGAGKTYSMEGPSILHCNKQKTGLVQRVVDLF	631			
Sbjct	61	DVYNFLAVPIVADAISGTINGIITYQGTGAGKTYSMEGPSILHCNKQKTGLVQRVVDLF	121			
Query	631	QSLQSSESMAMWSVKLSMVEIYLEKVRDLILSKDNLQIKESKTQGIYISGATEVSIQNS	811			
Sbjct	121	QSLQSSESMAMWSVKLSMVEIYLEKVRDLILSKDNLQIKESKTQGIYISGATEVSIQNS	181			
Query	811	SDALECLSEGIANRAVGETQMNLASSRSRSHCLYIFSVQQGSTSDERVRGGKIIILV	991			
Sbjct	181	SDALECLSEGIANRAVGETQMNLASSRSRSHCLYIFSVQQGSTSDERVRGGKIIILV	241			
Query	991	KVEKTGAEGRVLDEAKTINKSLSVLGNVNVNALTGKPNHVPYRDSKLTRILQDALGGNSR	1171			
Sbjct	241	KVEKTGAEGRVLDEAKTINKSLSVLGNVNVNALTGKPNHVPYRDSKLTRILQDALGGNSR	AALLCCCSPSASNAPESTLTVRGFTRTKLIKTPKSISPEVDSIKKPIPDHGQNDLDRD	301		
Query	1171	AALLCCCSPSASNAPESTLTVRGFTRTKLIKTPKSISPEVDSIKKPIPDHGQNDLDRD	AALLCCCSPSASNAPESTLTVRGFTRTKLIKTPKSISPEVDSIKKPIPDHGQNDLDRD			
Sbjct	301	AALLCCCSPSASNAPESTLTVRGFTRTKLIKTPKSISPEVDSIKKPIPDHGQNDLDRD				
Query	1351	ILN				
Sbjct	361	ILN				
Query	1531	etv				
Sbjct	421	ETV				

Query
S é q u e n c e
soumise

Related Information

S u b j e c t
s é q u e n c e
Banque

Comment evaluer l'alignement?

Score, evaluate etc.

[Download](#) [GenPept](#) [Graphics](#)

[▼ Next](#) [▲ Previous](#) [▲ Descriptions](#)

RecName: Full=Kinesin-like protein KIN-1; AltName: Full=AtKIN-1; AltName: Full=AtPSS1; AltName: Full=Pollen semi-sterility protein 1
Sequence ID: [sp|Q8GW44.1|KN1_ARATH](#) Length: 465 Number of Matches: 1

Range 1: 1 to 417 [GenPept](#) [Graphics](#)

			Identities	Positives	Gaps	Frame
Score	Expect	Method	257/420(61%)	330/420(78%)	8/420(1%)	+1
493 bits(1269)	1e-169	Compositional matrix adjust.				
Query	271	MSNVTCVRFRPLSHKE-RKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	447			
Sbjct	1	MSNVTCVRFRPLSHKE-RKTNGDKVCFKRLDSESFVFKDEREEDVIFSFD	60			

Related Information

Gene - associated gene details

Algorithm parameters

General Parameters

Max target sequences

100

Select the maximum number of aligned sequences to display



Expect threshold

10



Word size

6



Max matches in a query range

0



Scoring Parameters

Matrix

BLOSUM62



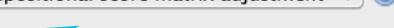
Gap Costs

Existence: 11 Extension: 1



Compositional adjustments

Conditional compositional score matrix adjustment



Filters and Masking

Filter

Low complexity regions



Mask

Mask for lookup table only



Mask lower case letters



Nombre max de séquences cibles

Seuil e-value

Taille amorce

Choix matrice de substitution

Score des gaps (existence extension) ?

BLAST

Search database UniProtKB/Swiss-Prot(swissprot) using Blastx (search protein)

Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

[RESTORE default search parameters](#)

2 alignements possibles, lequel choisir?

CGATGCAGCAGCAGCAGCATCG
||||||| |||||
CGATGC-----AGCATCG

CGATGCAGCAGCAGCAGCATCG
|| | | | | | | | | | | | | |
CG-TG-AGCA-CA--AT-G

Match= +1

Gap= -1

CGATGCAGCAGCAGCAGCATCG
||||||| |||||
CGATGC-----AGCATCG

CGATGCAGCAGCAGCAGCATCG
|| | | | | | | | | | | | | |
CG-TG-AGCA-CA--AT-G

$$(13 \times 1) + (6 \times -1) = 7$$

$$(13 \times 1) + (6 \times -1) = 7$$

Les 2 alignements ont le même score

CGATGCAGCAGCAGCAGCATCG
||||||| |||||
CGATGC-----AGCATCG

CGATGCAGCAGCAGCAGCATCG
|| | | | | | | | | | | | | |
CG-TG-AGCA-CA--AT-G

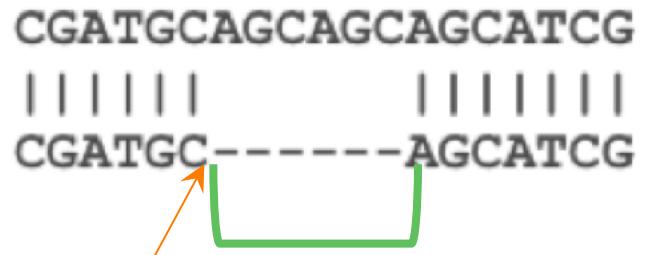
$$(13 \times 1) + (6 \times 1) = 7$$

$$(13 \times 1) + (6 \times 1) = 7$$

**Le 1er alignement est plus réaliste
(1 seul événement évolutif contre 3)**

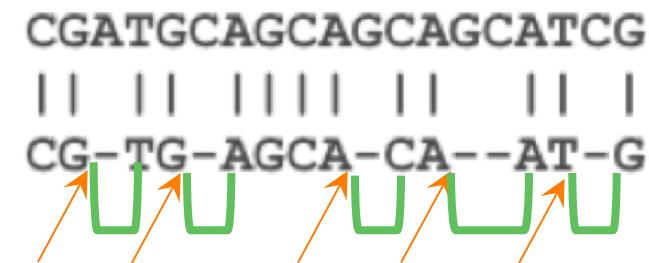
Match= +1
Gap= -1

Insertion / délétion : 2 paramètres
Ouverture gap -10
Extension gap -1



Ouverture gap Extension gap

$$-10 + 6 \times -1 + (13 \times 1) = -3$$



$$(-10 \times 5) - (6 \times 1) + (13 \times 1) = -43$$

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

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

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

Score = 46, Evalue = 4×10^{-4}

On s'attend à trouver en moyenne 0.0004 alignements de score 46 purement par hasard

Si je blaste 2500 séquences aléatoires -> 1 alignement

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

Score = 46, Evalue = 4×10^{-4}

On s'attend à trouver en moyenne 0.0004 alignements de score 46 purement par hasard

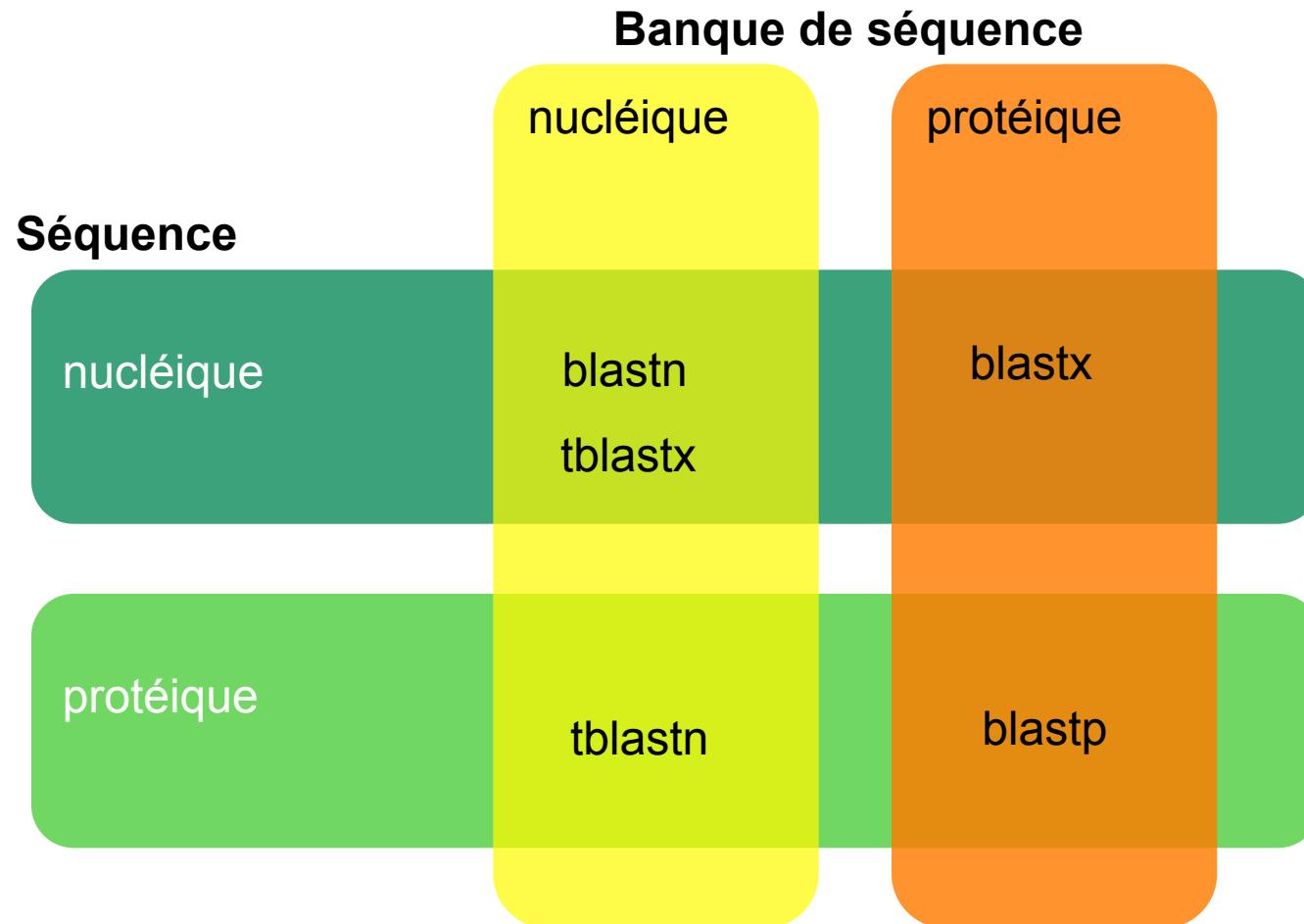
Si je blaste 2500 séquences aléatoires -> 1 alignement

Score = 267, Evalue = 1×10^{-70}

Il faut aligner 1×10^{-70} séquences aléatoires avant de tomber au hasard sur un alignement de cette qualité



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



Les logiciels

Ne les croyez pas systématiquement !!!

- Parfois diminution de la qualité des résultats au profit de la rapidité
- Recherche d'une solution parmi un ensemble infini de possibilités
- Ce n'est pas toujours la solution la meilleure qui est trouvée

Les banques de données ***Ne les croyez pas systématiquement !!!***

- > Les données ne sont pas toujours fiables ou à jour.
- > Différence entre réalité mathématique et réalité biologique

Les ordinateurs ne font pas de biologie, ils calculent . . . vite !



Merci!



Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International: <http://creativecommons.org/licenses/by-nc-sa/4.0/>

