

Introduction à la bioinformatique appliquée à la génomique

Christine Tranchant-Dubreuil

IRD

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

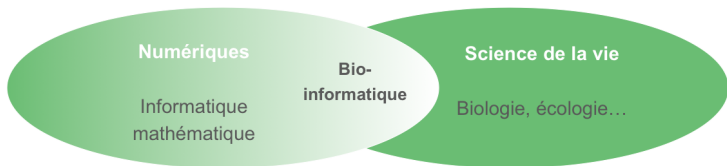


Pourquoi la bioinformatique ?

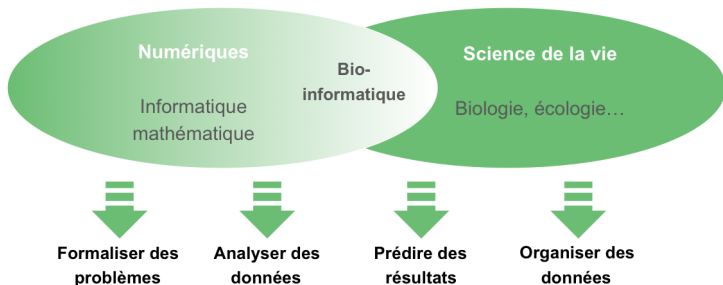
- Un nombre croissant de projets de séquençage
- Une masse de données génomiques publiques en croissance constante : banques de données nucléiques (ESTs, BACs), protéiques, génome etc.
- De nombreuses ressources web disponibles :
 - Définition *in silico* de marqueurs de type SSR, SNP, élément transposable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

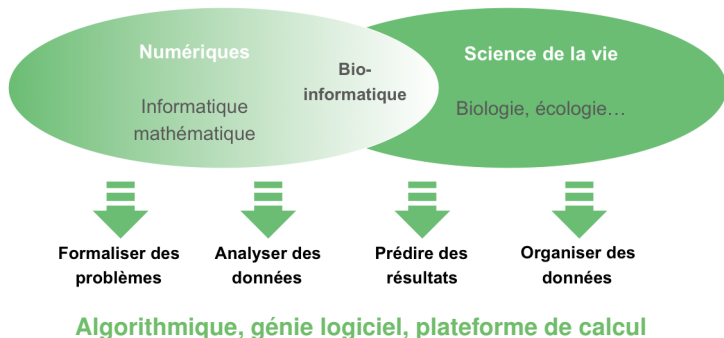
Qu'est ce que la bioinformatique ?



Qu'est ce que la bioinformatique ?



Qu'est ce que la bioinformatique ?



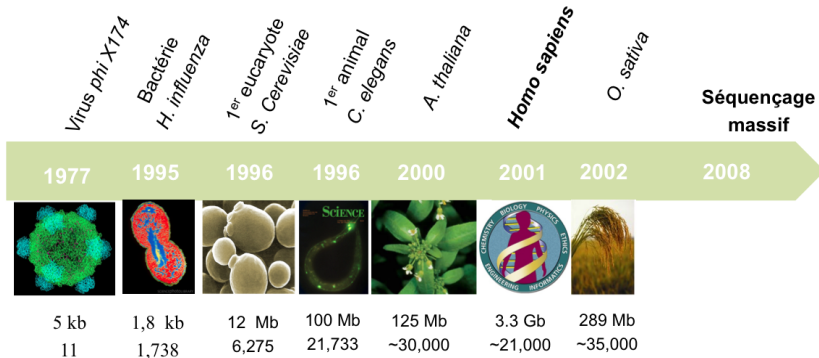
Génomique et bioinformatique ?

In silico vs in papyro ?

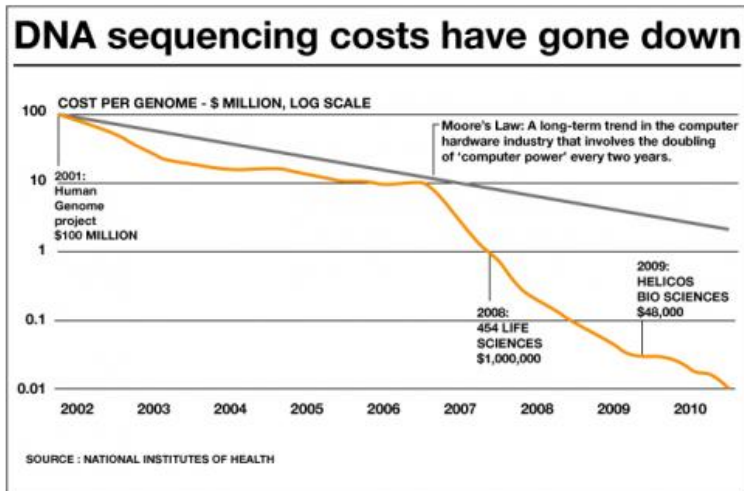
- **Génomique** : Etude des génomes et de l'ensemble de leurs gènes (structure, expression et régulation des gènes, évolution et dynamique)
- **Bioinformatique** : Approche *in silico* de la biologie indispensable pour :
 - Stocker et gérer les données de séquençage des génomes
 - Leur donner un sens

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

Evolution du séquençage des génomes



Evolution du séquençage des génomes



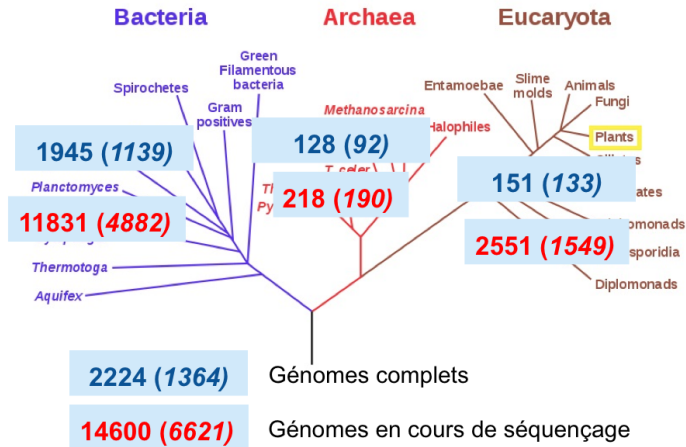
Révolution des technologies de séquençage



<i>1990</i>	<i>Actuellement</i>	<i>Défi</i>
Génome humain 4 milliards \$ (13 ans)	30000 \$ (quelques semaines)	Un génome humain 1000\$

Une explosion des projets de séquençage et des données génomiques

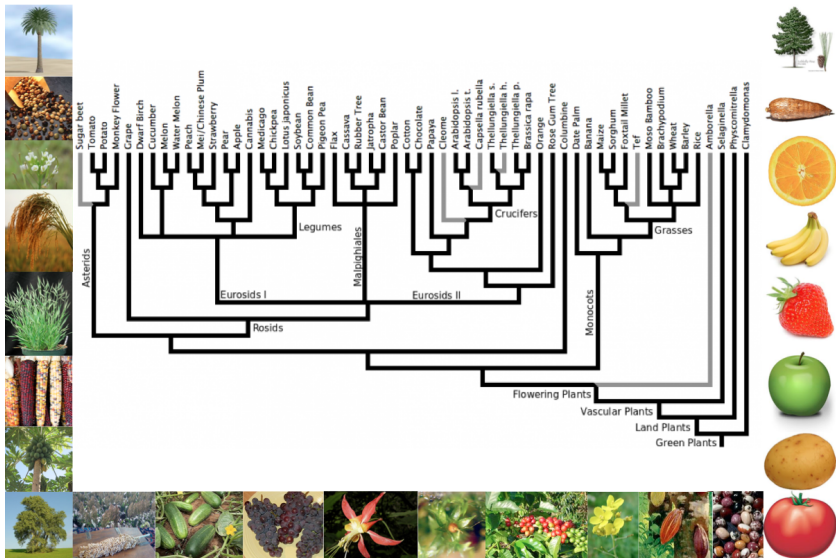
Bilan des projets génomes en 2012 (2010)



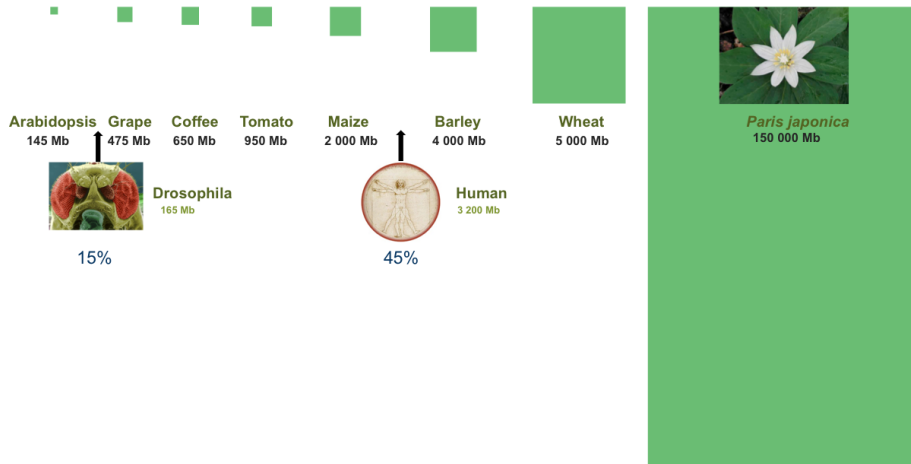
<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>

<http://www.genomesonline.org>

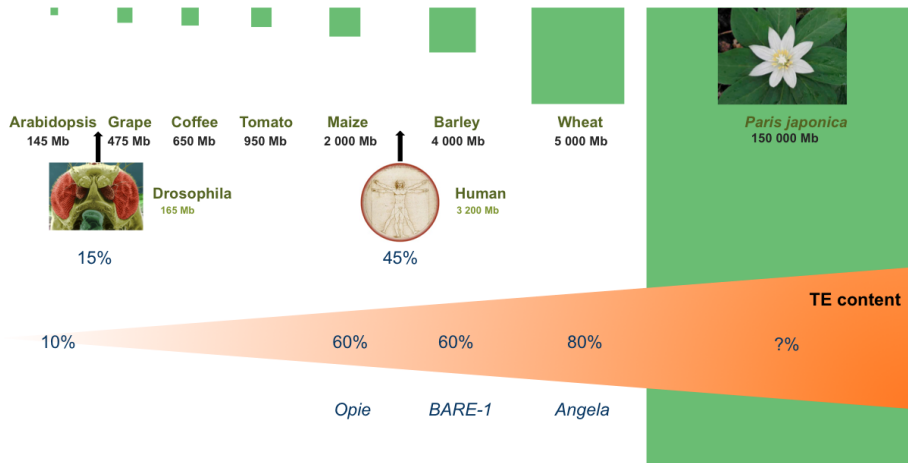
Arbre phylogénétique des génomes de plantes séquencés



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/>

The screenshot shows the homepage of the Genomes Online Database (GOLD). At the top left is the GOLD logo, and at the top right is the JGI logo (Joint Genome Institute, DOE Joint Genome Institute, US Department of Energy, Office of Science). Below the logos is a blue navigation bar with "Genomes Online Database" and "Home". The main heading reads "Welcome to the Genomes OnLine Database". A paragraph below states: "GOLD: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world." The page is divided into three columns: "Metagenomes", "Isolate Genomes", and "Genome Distribution".

Metagenomes	Isolate Genomes	Genome Distribution
Classification <ul style="list-style-type: none">• Studies: 409• Samples: 3510	<ul style="list-style-type: none">• Complete Projects: 7403• Incomplete Projects: 24913• Targeted Projects: 1566	<ul style="list-style-type: none">• Project Type• Sequencing Status• Phylogenetic

Left sidebar menu: Home, Genome Map, Genome Earth, Search, News, Statistics, Team, Reference.

Bottom right navigation icons: Home, Back, Forward, Stop, Refresh, Search, Print, etc.

Un peu de pratique... TP !

Comparatif des technologies NGS

<i>1990</i>	<i>Actuellement</i>	<i>Défi</i>
Génome humain 4 milliards \$ (13 ans)	30000 \$ (quelques semaines)	Un génome humain 1000\$

**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-benchttop-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 nécessaire
- **Amplifier l'ADN selon les technologies**

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

Les différentes étapes du projet de séquençage

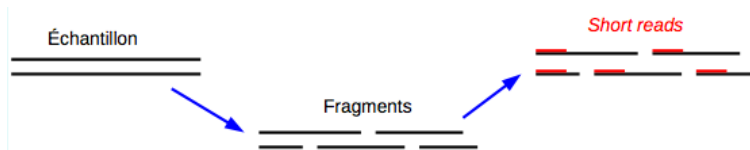
Lors du séquençage

- 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

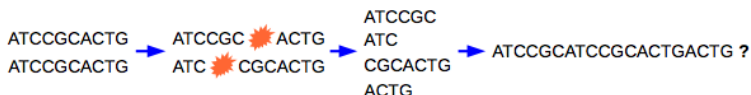


- Chaque séquence est présente en plusieurs exemplaires dans l'échantillon
en recoupant les reads, il est possible d'obtenir la séquence complète

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

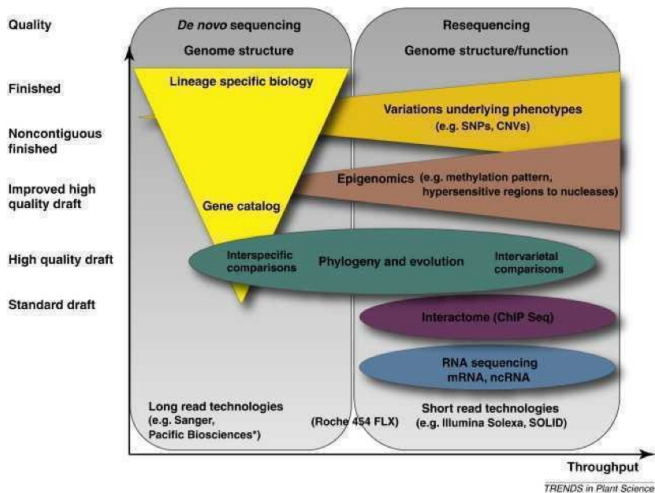
Les différentes étapes du projet de séquençage

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 ?

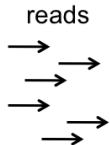


PHASE TWO: INTERPRETATION

SEIDMAN the New Yorker

Focus on Quality





assemblage



Mapping contre
une référence





Construct clone map and select mapped clones

AGTTCGTAACCTA	TGGCAATTGTAGA	CGATCGATGACFA
ATTGGACTTCGGA	TAACTCTGCATGCT	CAGCTAGCGGTGAT
CGATCGATGACTG	TGATCGATGTA	ATGCTGACTGTAG
CTTGATCGATGTA	GGATCTTACAAGT	ATAAAGTGCCTTG
ACTGGGATCCTAC	GGATTAATAAACCA	CGAGCGTTGCCAG
TGCGGTATAGCCC	AACGTTAGATCGA	ATCGATGTACTGG
AATCGATATCGAT	TAGCACATCGCGT	ATCTTACAAGTAA
ATACAGCTTCTAT	ATAGCCCGTAGAT	CGTTAGATCGATA
TAGATCGATGAAT	CGTGTATCGATAT	GCACATCCCGTAT

Generate several thousand sequence reads per clone



Assemble

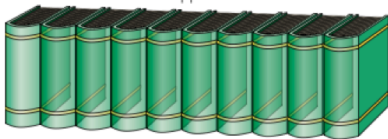
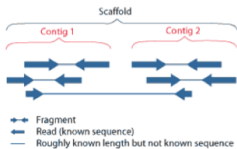
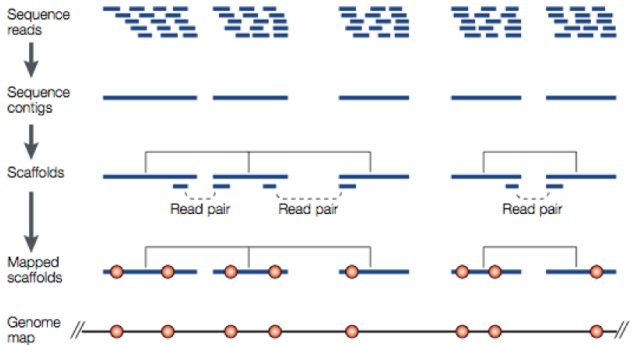


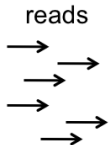
Generate tens of millions of sequence reads



Assemble



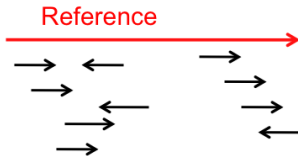




assemblage



Mapping contre
une référence



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

CAACTAGCAGCTAACCAGCAGAGAAAAATTAATGAGATTGAAAAAGCTGCAAGCACAAGAAAAATTCGTACAAAGCTCCATCC
AAGAATTGAGAGAAGAACAAAGTCAGCTCACTAGCTCCATCACTGCCTAAATAATAAAATGATAAGCTAATGGAAGAGCCAAA
GGAAAGATCAATCTGGACCTTCATCTTCAAAACCTTCAACCTAGACTGACCTTAGACTAAACCTAACCTAAGCTTAGACTGAAACC
CATGTTTTCAAATATAATCTTGACTAAGCTGAGTTTCATGATTATTTGCCACACTTGATGTACAGATACCAATTGATAACAATTG
CAATCTCTGATTTATCAAATGACTTTATGCAATGAATGATTACCTTTCTTGAACATATGCTTTCATTGGTTTATATACTGCTCTG
ATACATATTGCATCTGATACAAATCAAATCTGTAATGGTATCAAATCTTAATGTTGGTATCAGATATATTTTTAACTTCTCTCT
TTTGTGATGACAAAAAGGGGGAATATAAAAAAGAATCTCATCTGAGAAATGAGAAATATATGAAAAAGAAATCTAATCTGA
GAAATAAGAAATGTCAAAGAATTCAGAATACAAATCTCTGCACTAAAGCTAATATATAGGGGGAGATTTAAATATGATAAAGAAC
AAGAACTCTGCATAATTCGAATGTTTGAATGATAGGGGGGAGACCTTTCTCCATAATCTGTTGAAATTATGCATATATCTCTGAAC
GTCATATTTGCTCTGAACCTCATTATAGAATTCTAATAATTGCTGTGAAATTTATGTAATTTATTTAACTCAATGGTTTTGTCAT
CATAAAAAAATGGGGAGATTGTCAACCTAAAGGATGAATTTNN
NNNNNNNNNNNNNNNNNNNTGTCTGAAAACCAACTGGGTATTTGTGATCGCAGAAAAACAAAAGTTTTCGATTGGTTGCC
GAGAAAACCAACTGAATTTTGATTGTGAGCCCGAAAACAATCAGGCTGTAATCTGCTGGGTATAGTGAATCTCAAGCTAGG
CTTGAGGAGTGGACGTAGGTGCTGGGAGTGCATCGAACCACTATAAATCTTGGTGTGTTGATTGTGCTTCTTCTCTCTCTCT
CTCTGCATATCTGACATTTCTCATACTTTATTCACTGCTTATTGTACATTTCTATTTCCGCTGCTCAATCTTTAAAATGAAAGT
AACTCATACTCTCTCACGTTTTAACTTTAACTATTTTTAAAGACACCCAAATTCACCCCCCTCTGGGTGCTCAACTCTGGGCAA
CAATGAGCAATACCAATATAGTATTGAACGATACATATATACTTCTATTATAAGCTAAAATGAATGTGGAGAGTGATAAACATATTT
CCAATACAAAGGCAAATATAGTTTTAGCCCATGAGCATTCTGGTTGATGATGATCATTGTTGTTATATCTCCTGGATTTCATGG
TACATTCGTTGCTTTACCTTTCTATTGGCCAAAATCTCCTCCATTTGTTCTATAATTTGGACTAAAAATGAGTTAGATTTAATCTG
ACCGAATCTATTTCCATATCTGAAGTGTATCAGATCGAACACAAATTTGATATCCGAATTGATGTGGATCCGAATTTGGATTTA
AATACTTTTCGAATCCAATCTAGATATCCAGTAAAAAAAATTCAAAATAAACTGCATAACCTAGGATTCCAACCTAGAGATATA
GTTTGAATGGTAATGTCTTAAACACCAAGATACAATATATTCCTATCTTATTAATGAGTTATTTACTTGTATATGATTTCAGG
TCATTAATATCATATAATGTATAGCATGATATTTATATTTAATCACTATTTTTATAAAAATCTTCTCTTTGATAAAAATGAAAATGA
ACATACCTTTATTTATTTATAATATGTTTGAACCTTTATTGTAAAATATTAATAAATAAACTTCTATTTGTTCTTATATGCATATAAT
TTTGTCTACAAGCTTTATAAGAATATATAATCACAATTTTTATTATGCTATAATAAAAGTAATAATAAATGATACAAAACATG
CTAGTATTTTTATGTTGATTTCTTATCTTAAATGATTTTTATTAAATTTAATAAAAAATCTGGTCAACCAATGACCTAATCCTTGAC
AAGGTCAGTCTTCATGTCAGTTTAATAACCATATCTTAGATCCTTGCTTACTTGCATCGTGTGCAATTTGCATAGATAAAAAATA
TATCCGATTTATATTTGATGTTTAAAAACAATATGCATATGCTTAATATCCGATCTATATCCGATCTATTTAGAACACAATGCAA
CTAATTTGGTATTCATTTATATCCATATGGATCTATATTTGTAATAAATATGGATACAAATATGGATGTACCAGTATTCGATCAATAT
CCAATTTGTTTTACTCCATTAACATCAGATAGTGGCAGATATCATGGACTTATACTTTTGTCTGCAAGCTCTCTTTTAGTGG
AGGCTTGCCAATGCTTTAGTTGGGACATCATAGGATTCTATGGATGGTGGGGCTTATATTTTTCTTGGACCCTGCCTAATA



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

**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

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 ?

**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

**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 !!

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

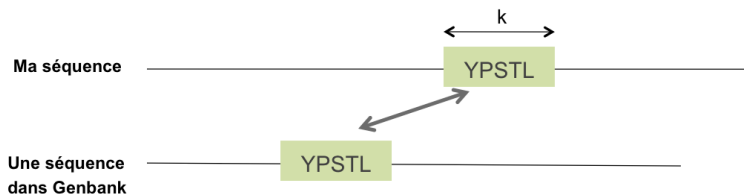
Solution 1

- 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

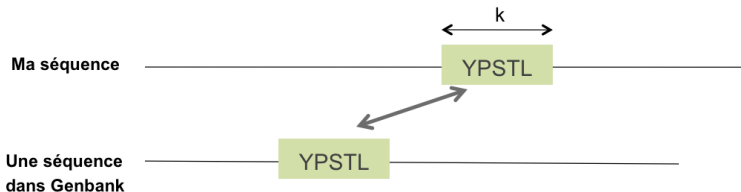
- Faire une pré-sélection sur les séquences puis aligner exactement avec SW
- Avantage : Rapide et efficace
- Inconvénient : Risque de passer à côté de la perle !
- Algorithme : Smith-Waterman
- Alignement local
- BLAST
- Heuristique

**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 ?

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

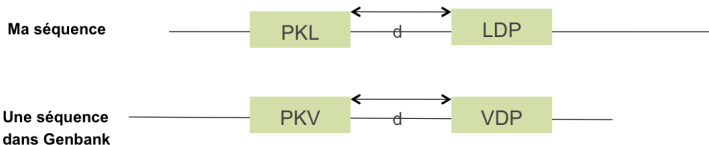


Ne retenir que les séquences partageant au moins un mot de longueur k avec ma séquence



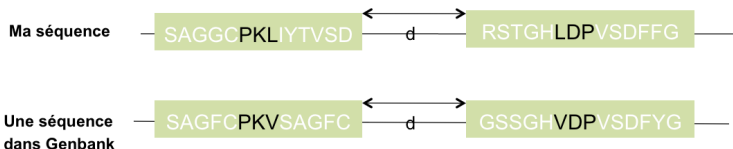
BLAST indexe les séquences et détermine, pour tous les mots de longueur k , la liste des séquences qui contiennent ce mot

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



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

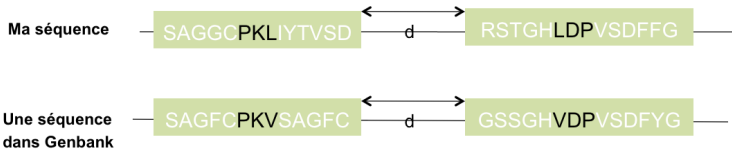
**Quelle est la similarité entre ces 2 séquences?
Est ce que ces séquences sont homologues?**



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

The screenshot shows the NCBI BLAST web interface. Three red annotations with arrows point to specific parts of the form:

- 1. requête (votre séquence)**: Points to the "Enter accession number, gi, or FASTA sequence" input field.
- 2. portée (à qui vous voulez la comparer)**: Points to the "Database" dropdown menu, which is currently set to "Non-redundant protein sequences (nr)".
- 3. et hop !**: Points to the "BLAST" button at the bottom of the form.

The interface includes fields for "Enter Query Sequence", "Job Title", "Choose Search Set" (Database, Organism, Enter Query), and "Program Selection" (Algorithm). The "BLAST" button is labeled "Search database nr using Blastp (protein-protein BLAST)".

Annotation structurale : Blast

The screenshot shows the NCBI BLAST search page. Three red annotations with arrows point to specific parts of the form:

- 1. requête (votre séquence)**: Points to the "Enter accession number, gi, or FASTA sequence" input field.
- 2. portée (à qui vous voulez la comparer)**: Points to the "Database" dropdown menu, which is currently set to "Non-redundant protein sequences (nr)".
- 3. et hop !**: Points to the "BLAST" button at the bottom of the form.

The form includes fields for "Query subrange", "Job Title", "Align two or more sequences", "Organism", "Enter Query", and "Program Selection" (with options for blastp, PSI-BLAST, and PSI-BLAST).

1. récapitulatif de la requête

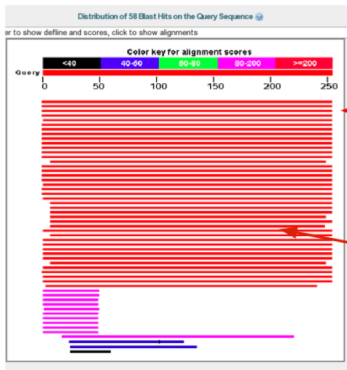
quelle séquence a été soumise ("query") ;
identifiant, longueur, type

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

sp|P04156|PRIO_HUMAN Major prion protein OS=Homo... quelle banque de donnée est interrogée ?

Query ID	Id 46010	Database Name	swissprot
Description	sp P04156 PRIO_HUMAN Major prion protein OS=Homo sapiens GN=PRNP PE=1 SV=1	Description	Non-redundant SwissProt sequences
Molecule type	amino acid	Program	BLASTP 2.2.21+ Citation
Query Length	253		quel programme est utilisé ?

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#)



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] >gil94730359[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 glucof	842	842	99%	0.0
P14804.3	RecName: Full=Glucoamylase; AltName: Full=1,4-alpha-D-glucan glucohy	631	631	95%	0.0
Q23045.1	RecName: Full=Glucoamylase P; AltName: Full=1,4-alpha-D-glucan glucc	560	560	99%	0.0
Q60087.1	RecName: Full=Probable glucoamylase; AltName: Full=1,4-alpha-D-glucan	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
P04095.2	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
Q30565.1	RecName: Full=Cyclomaltoextrin 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=Cyclomaltoextrin glucanotransferase; AltName: Full=Cyc	71.2	71.2	14%	1e-11

seq|P27177_2|PR10_CHECK RecName: Full=Major prion protein homolog; AltName: Full=PR-1P; AltName: Full=Acetylcholine receptor-inducing activity; Short=ARIA; AltName: Full=65-21 protein; Flags: Precursor Length=273

GENE ID: 396452 PRNP | prion protein (p27-30) (Creutzfeldt-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) [Gallus gallus] (19 or fewer PubMed links)

Score = 80.9 bits (198), Expect = 6e-15, Method: Compositional matrix adjust. Identities = 94/230 (40%), Positives = 120/230 (52%), Gaps = 33/230 (14%)

```
Query 18 DLCLL...KKRKPQGRNTGSKYFDGSSPGGRKTFYDGGGGRGPPDGGGPPGGGG 74
Sbjct 20 DVALSKKGGKSGGGGGAGSHRQPSYPROPDPYHPNPQYHPNPQY--PHNPQY- 76
Query 75 QPHGGGGG---PHG--GGWG-----GGGTHQWPKPK-TKTNQWAGAAAAA 120
Sbjct 77 -PHNPQYQNPQYHPNPQYPGGGGQYNPSSGGSYHQ--KPNRPPKTNFKHAGAAAAA 133
Query 121 VVGLGGYRLGASRPIIHFGSDYEDRYRENBRYPNGVYRPFDEYSNONNFVHDCV 180
Sbjct 134 VVGLGGY *G HS HF S E R+ E RYPN+VYR Q+ FV DC 193
Query 181 NI I KQHTVTTTK-----GENFTEIDV-KMREKVVQKICITOVER 220
Sbjct 194 NI I TVEYSI GPAAXKXITSEAVAAANQTEVEKENVKVTKVIRENCVQYRE 243
```

seq|A415K2_1|IF2_DESBM RecName: Full=Translation initiation factor IF-2 Length=95

GENE ID: 4955422 Dred_1957 | translation initiation factor IF-2 (Neospora caninum radicans H-1)

Score = 46.6 bits (109), Expect = 2e-04, Method: Composition-based stats. Identities = 39/103 (37%), Positives = 51/103 (49%), Gaps = 32/103 (31%)

```
Query 29 GGNITGGSKYF-----GGGSPGGRYFPQG-----GGGQPHG---GGGG 67
Sbjct 165 GG GG P GGG P G+R PGG GG G+P+G GG G+ 222
Query 68 PHG---GGGQPHG---GGGQPHGGGGGGTHSQWPK 182
Sbjct 223 PYGRPQGGGQRYGRPQGGGQRYGR-PGGGQSRPYGRP 264
```

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

Algorithm parameters

General Parameters

Max target sequences: 100 *nombre max. de séquences cibles*
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10 *seuil sur l'E-value*

Word size: 3 *taille de l'amorce*

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62 *choix de la matrice de substitution*

Gap Costs: Existence: 11 Extension: 1 *score des gaps pourquoi y a-t-il 2 paramètres ???
- Existence
- Extension*

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

Annotation structurale : Blast

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



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

Match = +1

Gap = -1

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

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



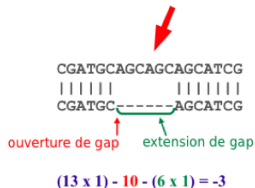
```
CGATGCAGCAGCAGCATCG
|| || |||| | | | |
CG-TG-AGCA-CA--AT-G
```

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

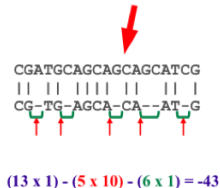
- les 2 alignements ont le même score

Annotation structurale : Blast

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



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



- **insertion/délétion: 2 paramètres**

- ouverture de gap (par ex -10)
- extension de gap (par ex -1)

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 matches similaires à celui obtenu soient trouvés simplement par hasard,

$s = 46 \text{ \> Evalue} = 4e-4$: 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)

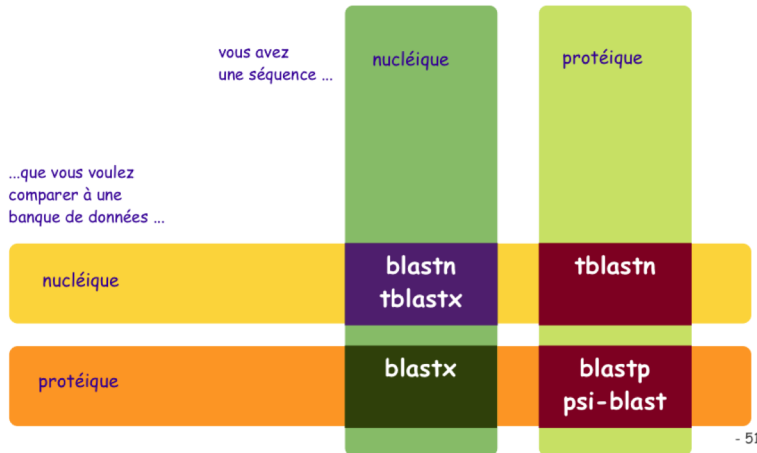
$s = 267 \text{ \> Evalue} = 1e-70$: il faut que je blaste $1e70$ séquences aléatoires avant de tomber au hasard sur un alignement de cette qualité ...

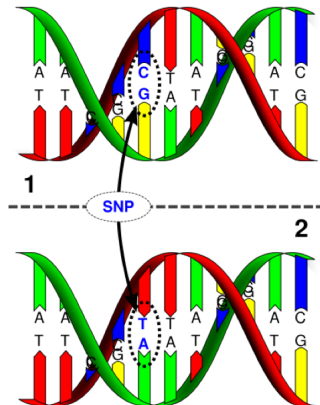
Annotation structurale : Blast



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

Annotation structurale : Blast





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 SVTYPE=DEL;END=300

The image shows two overlapping screenshots of the 1000 Genomes Project website. The background screenshot is the main project page, titled "1001 Genomes: A Catalog of Arabidopsis thaliana Genetic Variation". It features a navigation bar with links for Home, Collaborators, Accessions, Tools, Software, Data Center, Gallery, About, and Help desk. A central banner reads "Welcome to the 1001 Genomes Project" with buttons for "Check" and "Browse". A "Download" box on the right encourages users to use the Data Center to download project-related SNPs, indels, SVs, and genome sequences. A "Links" section lists resources like GBrowse, Assemblies project, POLYMORPH, NCBI SRA Genomes Project, and a map resource for 1001 Genomes. A "News" section mentions an October 22, 2010 update regarding the VAFDBM database.

The foreground screenshot is a dark-themed version of the website, titled "1000 Genomes: A Deep Catalog of Human Genetic Variation". It has a navigation bar with links for Home, About, Data, Analysis, Participants, Contact, Browser, Wiki, and FTP search. A search bar is present. The main content area is divided into "LATEST ANNOUNCEMENTS" and "LINKS".

LATEST ANNOUNCEMENTS

WEDNESDAY FEBRUARY 16, 2011
February 2011 Data Update
[Full Project Indel Release](#)
Indel calls from [Dindel](#). These calls are based on 629 individuals from the 20100804 sequence and alignment release of the 1000 genomes project. This release is based on the GRCh37 assembly of the human genome and are released in the format VCF 4.0
Data access links: [EBI/NCBI](#)
Link to additional information: [README file](#)

THURSDAY DECEMBER 16, 2010
December 2010 Data Update
[Full Project Genotype Release](#)

LINKS


- All Project Announcements
- Sample and Project Information
- Media Archive
- Download the 1000 Genomes Pilot Paper




Outil collaboratif

The screenshot displays the Galaxy web interface. On the left is a sidebar with a 'Tools' menu listing various analysis tools such as GAT, Bowtie, and BLAST. The main content area features a central banner for the 'Galaxy Community Conference 2012' with the text 'Registration is now open! Abstract deadline: April 16'. Below the banner is a 'Live Quizzes' section with several tool icons. At the bottom, there is a 'galaxyproject' logo and a 'Sponsored Report' section with text about the Galaxy team and their work at various institutions.

The iPlant Collaborative

The iPlant Collaborative develops cyberinfrastructure and computational tools to solve Grand Challenges in plant science



CHALLENGE	DISCOVER	LEARN	CONNECT
<p>iPlant Genotype to Phenotype (iPG2P)</p> <p>Mapping the links between genotypes and phenotypes</p> 	<p>Discovery Environment</p> <p>Access iPlant tools through a single user-friendly interface</p> <p>MORE...</p>	<p>Upcoming Events</p> <ul style="list-style-type: none">• Biology of Genomes May 08 2012 - May 12 2012• iPlant Tools and Services Workshop @ University of Arkansas, Little Rock May 17 2012 - May 18 2012• iPlant Tools and Services Workshop @ Purdue University May 21 2012 - May 22 2012 <p>MORE...</p>	<p>People at iPlant</p> <p>Community driven science</p> 
<p>iPlant Tree of Life (iPToL)</p> <p>Understanding the phylogenetic relationships between all plant life</p> 	<p>DNA Subway</p> <p>An educator-tailored interface for bringing iPlant to the classroom</p> <p>MORE...</p>	<p>the iPlant^{Leaflet}</p>	

The screenshot displays the iPlant Collaborative Discovery Environment interface. The top header includes the iPlant Collaborative logo and the text "Discovery Environment", along with a user ID "rg1971" and links for "Help" and "Notifications". On the left sidebar, there are three main icons: "Data", "Analyses", and "Apps".

The main content area is divided into two overlapping windows:

- Analyses Overview:** This window shows a search bar with the text "Filter by Name or App" and a "View Output(s)" button. Below the search bar, there is a section for "Apps" with a "No items to display" message.
- Apps Panel:** This panel is divided into two sections:
 - Categories:** A tree view showing various application categories such as "Workspaces (2)", "Applications under development", "Favorite Applications (2)", "Public Applications (184)", "Bio (43)", "NGS (38)", "Aligners (4)", "QC and Processing (6)", "Assembly Annotation (4)", "Transcriptome Profiling (8)", "ChIPseq (1)", "Utilities (4)", "Variant Identification (1)", "Assemblers (3)", "SAMTools (7)", "GTL and GWAS (12)", "Data Sources (2)", "Phylogenetics (2)", "Tree Building (1)", "Comparative Methods (4)", and "Evolutionary Models (2)".
 - Aligners Table:** A table listing specific aligner applications. The table has columns for "Name", "Integrated by", "Published on", and "Rating".

Name	Integrated by	Published on	Rating
BWA (Paired-End Illumina Reads)	Matthew Vaughn		★★★★★
BWA (Single-End Illumina Reads)	Matthew Vaughn		★★★★★
SOAP2 (Single-End Illumina Reads)	Matthew Vaughn		★★★★★
SOAP2 (Paired-End Illumina Reads)	Matthew Vaughn		★★★★★

Un site utile pour les analyses NGS

SEQanswers
the next generation sequencing community

Home | FAQ | Community | Calendar | New Posts | Search | Quick Links | Log Out

New Posts

Last Post	Replies	Views	Forum
How to identify the configuration of the genome... 10/24/2012 11:42 AM By genomax 6	2	35	Bioinformatics
(NEW) pathogen 1.0.141 10/24/2012 11:42 AM By genomax 6	1	47	Bioinformatics
Data center ready in FASTA format 10/24/2012 11:42 AM By genomax 6	1	43	Bioinformatics
How to find fusion genes using bioinformatics... 10/24/2012 11:42 AM By genomax 6	13	1,370	Bioinformatics
Panel and library 300-mpx 10/24/2012 11:42 AM By genomax 6	3	124	Bioinformatics
Identifying contamination 10/24/2012 11:42 AM By genomax 6	11	1,060	Bioinformatics

New Posts

[Let's make the SEQanswers data more accessible](#)
10/24/2012 11:42 AM - By [genomax](#)

[Help make the wiki accessible](#)
10/24/2012 11:42 AM - By [genomax](#)

The SEQanswers wiki (or "database" for short) is a great help for users of the forum. It's a catalog of high-throughput sequencing tools. There is currently an effort to get the SEQanswers forum and the wiki published in the next NAR database issue. This is a **great opportunity** to get the wiki in shape. Some of the tool descriptions are just "dubs", but in order for these pages to be really helpful, we need just **two minutes of your time**.

We invite everyone to pitch... [Open topic](#)

17 Replies | 1,077 Views

New Posts

[Let's publish the Wiki!](#)
10/24/2012 11:42 AM - By [genomax](#)

There is now a great opportunity to publish [this](#) as a 'wiki-database' in the 2012 NAR annual database issue [1].

Below are the details from the previous email exchange with the editors of NAR. I've started a preliminary page for collaborative authoring of the paper here:

[Open topic](#)

138 Replies | 3,370 Views

New Posts

[Help! Read Address Deleted](#)
10/24/2012 11:42 PM - By [genomax](#)