H3ABioNet

معهد باستور تونس
Institut Pasteur de Tunis

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course: IBT

Genomics

Comparative Genomics

Learning Objectives

Session 3:
Comparative Genomics

- **Part 1: (intro)** Comparative Genomics: what is it?
- **Part 2:** Genomic Variation / Comparative Genomics: WWWH?
- **Part 3:** The input
- **Part 4:** The methods
- **Part 5:** The output

Learning Outcomes

Session 1:
Comparative Genomics

- Navigate through genomic ressources to:

  - Retrieve information on a specific gene (sequence, variants, orthologs…)

  - View and interpret genomic alignments

Part 1

(intro)
Comparative Genomics:
What is it?

Comparative Genomics: a definition

- Comparative genomics is based on the fact that a genomic variation is happening in all organisms.
- These changes affects several features in a genome (structure, organization, functions…).
- The changes could help monitor evolution between organisms (species…)

Comparative Genomics: a definition

- Comparative genomics deals with the process of comparing the sequences of **whole** or **parts of genomes.**

- Goal:
  - identify **similarities and differences** between **features** in these genomes
  - Identify **evolutionary relationship** between organisms

Comparative Genomics: a definition

- Comparative genomics : **features** + **evolution**

ORGANISM A1
GENOME A1
FEATURES A1

ORGANISM A2
GENOME A2
FEATURES A2

ORGANISM B
GENOME B
FEATURE B

Adapted from http://recettes-aymen.over-blog.fr/
http://quebueno.be/content/6

Part 2

Genomic Variation & Comparative Genomics: WWWH?

Genomic variation : WWWH

Genomic variation : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- **Why would a genome evolve ?**

→ Genomic plasticity allows an organism to:

- adapt to environmental changes
- Find the best evolution path
- Acquire virulence genes, enhanced pathogenicity
- Resistance to drugs
- Increase survival chances of members of a population
- …

Genomic variation : WWWH

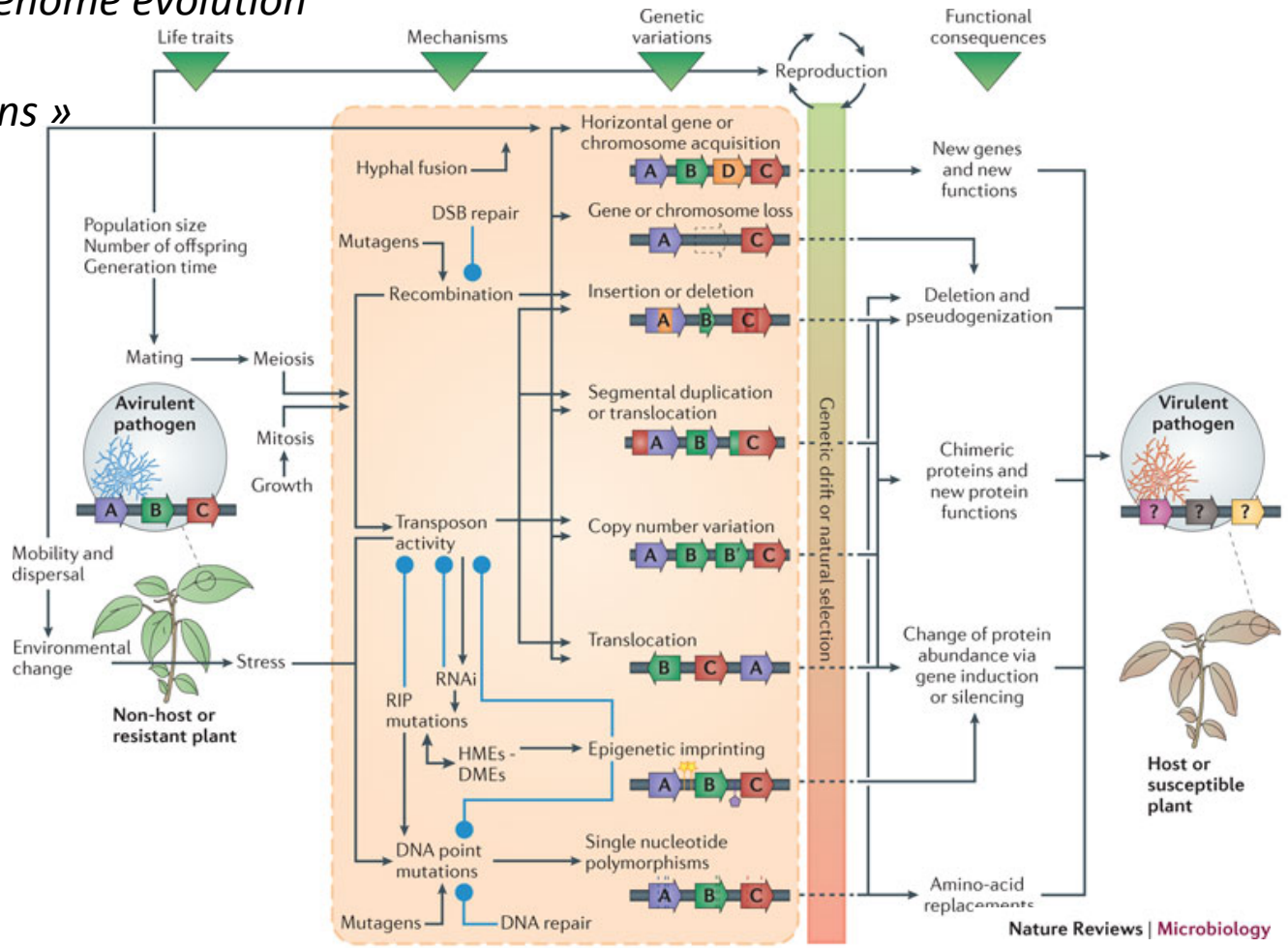
WHY ?

WHEN ?

WHAT ?

HOW ?

Example : *« Genome evolution in filamentous plant pathogens »*



(Raffaele & Kamoun, 2012)

Genomic variation : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- **Factors/events**

  - Gene transfer
  - Environmental pressure for selection
    - pH
    - temperature
    - host
    - pathogen
  - …

→ A genetic variation could occur in response to such factors / events

Genomic variation : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- **What could be affected ?**

  - Overall genomic sequence (re-arrangements)
  - DNA structure
  - Regulatory elements
  - Genes size, number, function, density
  - Nucleotide composition
  - …

Genomic variation : WWWH

WHY ?

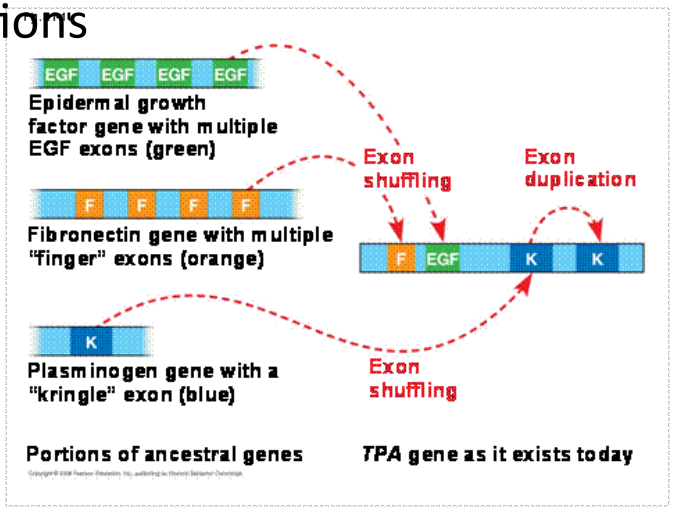
WHEN ?

WHAT ?

HOW ?

- **How could this happen ?**

  - Large genetic structural variations (duplication, recombination…)
  - Transposable elements (retrotransposons…)
  - Evolution of multigene families
  - Evolution of genes with novel functions
  - Exon shuffling
  - Tandem repeats modification
  - …



http://www.mrschamberlain.com/

Genomic variation : WWWH

WHY ?

WHEN ?

WHAT ?

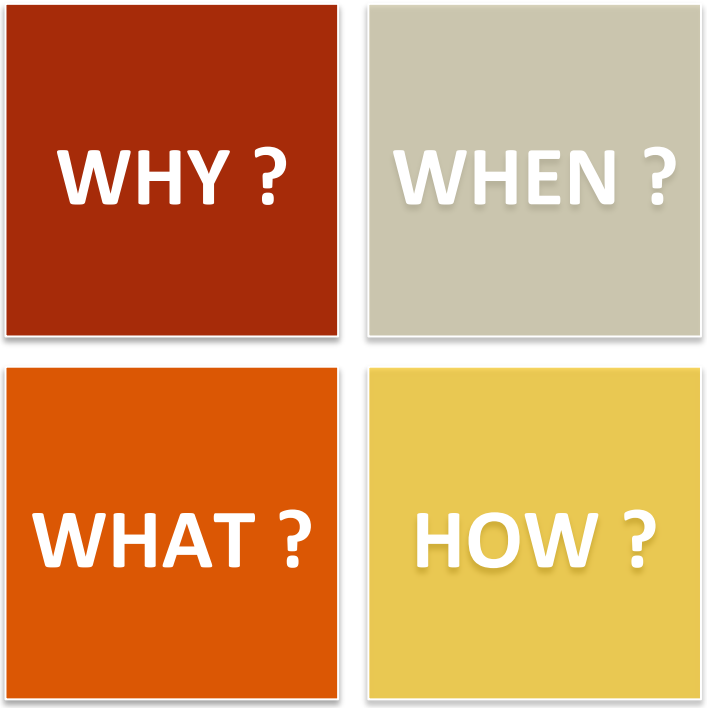
HOW ?

- **How to measure the changes in a genome?**

  - Sequence variation ← **Comparative Genomics**
    - between 2 genomes (1 reference)
    - between several genomes

  - Other variations (structure, folding…)

Comparative Genomics : WWWH



Comparative Genomics : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- **Why would we compare genomes ?**

  - Identify evolutionary history
  - Highlight synteny
  - Identify genomic rearrangements (large SV events…)
  - Study convergent evolution for some organisms (e.g. virus)
  - understand disease outbreak
  - Identify pathogenicity markers, drug targets
  - …

Comparative Genomics : WWWH

WHY ?

WHEN ?

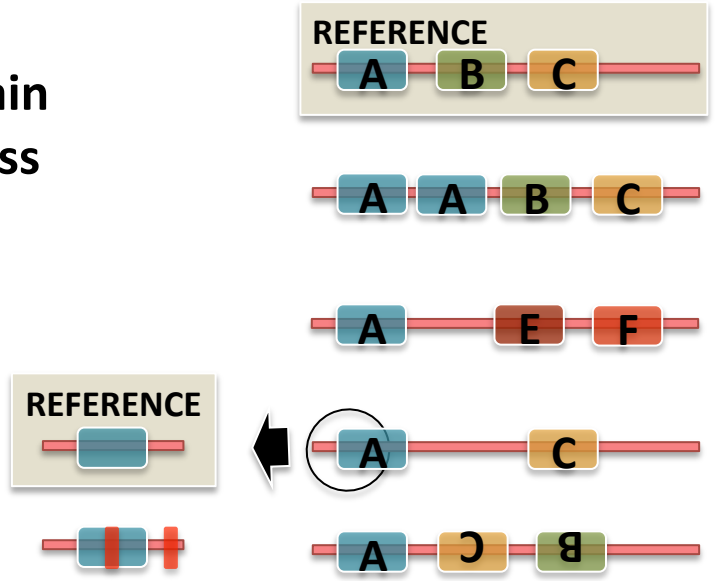
WHAT ?

HOW ?

Why would we compare “portions” of genomes ?

- Comparing smaller portions of a genome allows to zoom into regions of genomic re-arrangements

- Could be **genes**:
  - Screen for functional genes **gain**
  - Screen for functional genes **loss**
  - Gain of a **new function**
  - **Exons** (length, number…)
  - Conserved **pathways**
  - **Coding / non-coding**
- …



Comparative Genomics : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- **When do we need to use comparative genomics ?**

→ Establish genetic and evolutionary relationship between :

- Entire organisms
- Sequences

Comparative Genomics : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- What we generally compare are **features** of 1 or more genomes to features of a another **genome (reference)**
- A genome is complex and composed of different elements (regulatory, stuctural…)
- In fact, there are different types of DNA features that can be compared between 2 genomes:
  - **DNA sequences** (small, large, coding/non-coding)
  - **Genes** (nature, order…)
  - Regulatory elements
  - …

Comparative Genomics : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- Could be classified in:

  - Genome **structure**

  - Genome **function** (coding / non-coding)

  - Genome **evolution**

Comparative Genomics : WWWH

WHY ?

WHEN ?

WHAT ?

HOW ?

- Comparative genomics uses **Sequence Alignment**
- Comparative genomics is based on **Phylogeny** that relies on several key issues:

  - Several genomes are sequenced and available
  - Homology between genes (similar functions)
  - …

→Use **complex** model genomes to **infer knowledge** (**Annotation**: function…) to **unknown or less complex** genomes

Comparative Genomics : WWWH

WHY ?

WHEN ?

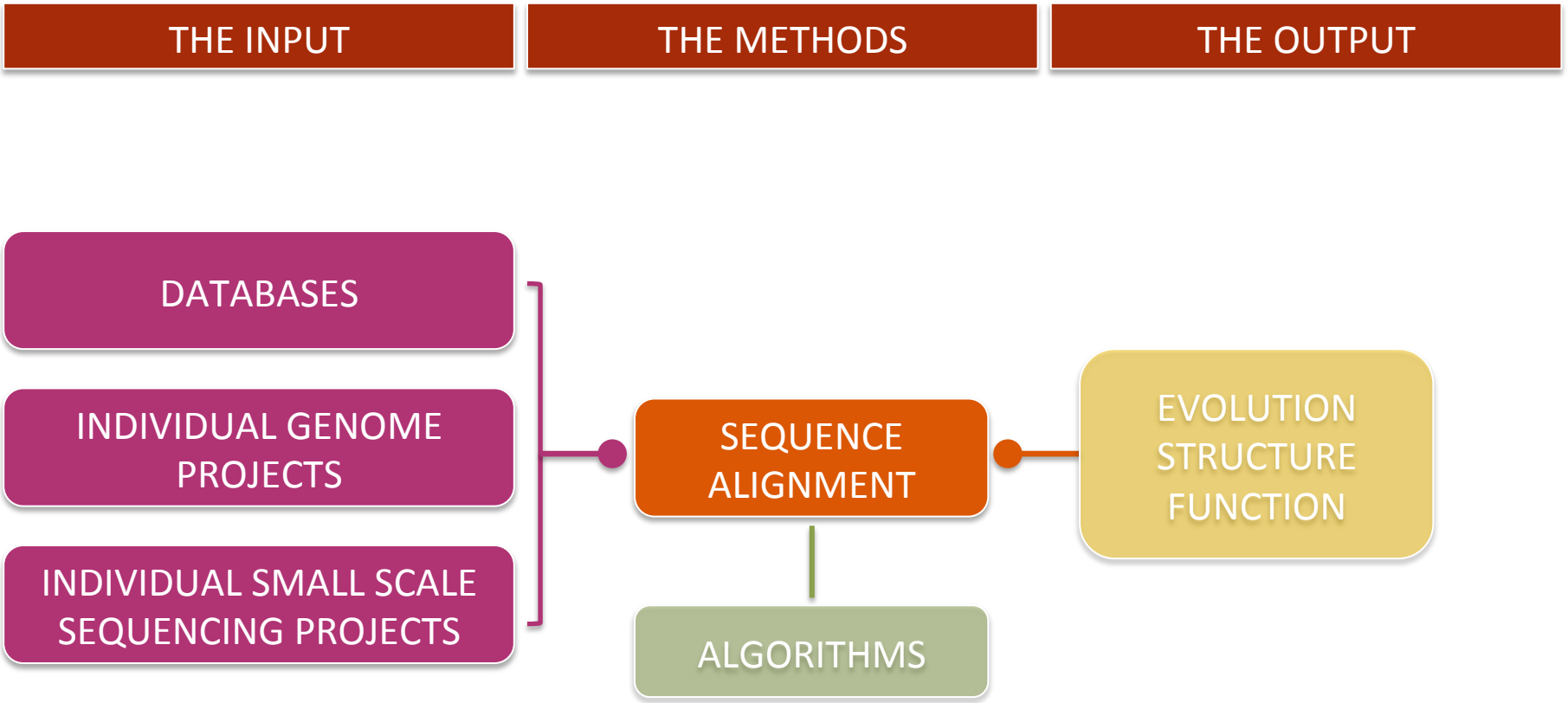
WHAT ?

HOW ?

- Algorithms/programs
- *in vitro:*

  - Fluorescence In Situ Hybridization (FISH)
  - Spectral Karyotyping (SKY) and Multiplex-FISH (M-FISH)
  - Comparative Genomic Hybridization

Keywords in Comparative Genomics



H3ABioNet

معهد باستور تونس
Institut Pasteur de Tunis

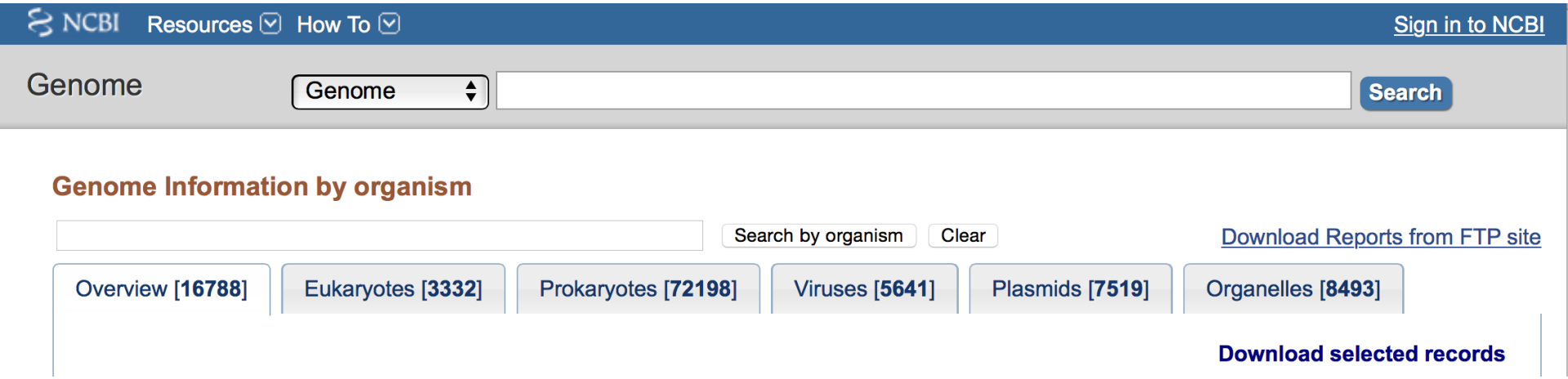
Part 3

Comparative Genomics: The input

Availability of genomes sequenced

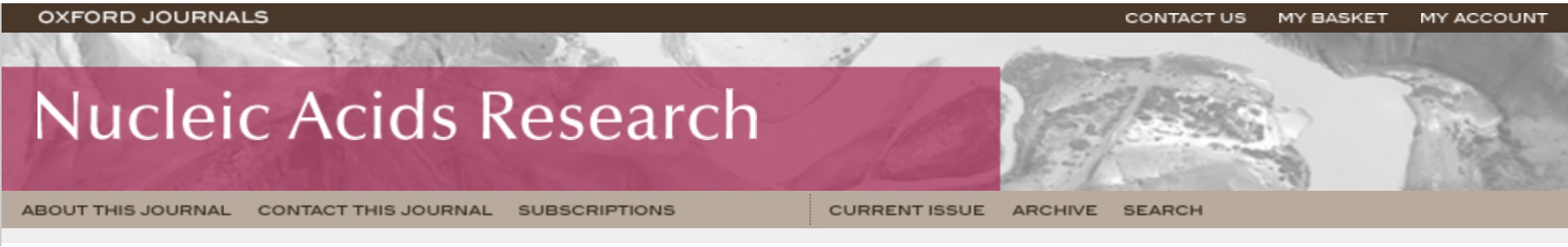
Perman, 2014: http://blogs.biomedcentral.com/

Availability of genomes sequenced



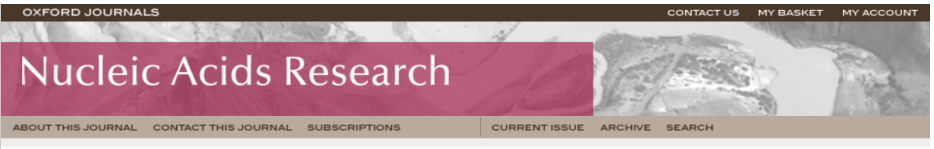
http://www.ncbi.nlm.nih.gov/genome/browse/

Availability of databases



https://nar.oxfordjournals.org/
http://database.oxfordjournals.org/

Availability of databases



Assembly: a resource for assembled genomes at NCBI

Paul A. Kitts*, Deanna M. Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G. Smith, Tatiana Tatusova, Charlie Xiang, Andrey Zherikov, Michael DiCuccio, Terence D. Murphy, Kim D. Pruitt and Avi Kimchi

+ Author Affiliations

+ Author Notes

*To whom correspondence should be addressed. Tel: +1 301 435 6091; Fax: +1 301 480 0109; Email: kitts@ncbi.nlm.nih.gov

Received September 12, 2015.
Revision received October 21, 2015.
Accepted October 29, 2015.

*« The **NCBI Assembly database** (**www.ncbi.nlm.nih.gov/assembly/**) provides stable **accessioning and data tracking** for genome assembly data (…)*

The Assembly database reports **metadata** such as assembly names, simple statistical reports of the assembly (number of contigs and scaffolds, contiguity metrics such as contig N50, total sequence length and total gap length) as well as the assembly update history. *The Assembly database also tracks the **relationship between an assembly submitted to the International Nucleotide Sequence Database Consortium (INSDC) and the assembly represented in the NCBI RefSeq project.** Users can find **assemblies of interest by querying the Assembly Resource directly or by browsing available assemblies for a particular organism.** Links in the Assembly Resource allow users to easily download sequence and annotations for current versions of genome assemblies from the NCBI genomes FTP site»*

Availability of databases

ANItools web: a web tool for fast genome comparison within multiple bacterial strains

Na Han[1,2], Yujun Qiang[1,2] and Wen Zhang[1,2,*]
+ Author Affiliations
*Corresponding author: Tel: (86)-010-61739446; Email: zhangwen@icdc.cn

Received November 25, 2015.
Revision received April 5, 2016.
Accepted April 28, 2016.

*« **Background:** Early classification of prokaryotes was based solely on phenotypic similarities, but modern prokaryote characterization has been strongly influenced by advances in genetic methods. **With the fast development of the sequencing technology, the ever increasing number of genomic sequences per species offers the possibility for developing distance determinations based on whole-genome information**. The average nucleotide identity **(ANI), calculated from pair-wise comparisons of all sequences shared between two given strains, has been proposed as the new metrics for bacterial species definition and classification.** **Results:** In this study, we developed the web version of ANItools (http://ani.mypathogen.cn/), which helps users directly get ANI values from online sources. A database covering ANI values of any two strains in a genus was also included (2773 strains, 1487 species and 668 genera) »*

http://database.oxfordjournals.org/

File formats

- **Different file formats may be accepted**
- **Blast (NCBI) accepts:**
  - Fasta sequence
  - simple sequence
  - Accession Number
  - Local file from disk (megablast)
- **Whole sequence (or subsequence)**

FASTA

```
>gi|343488507:5001-16300 Homo sapiens CD74 molecule (CD74), RefSeqGene on
chromosome 5
CTGCCTGGGGAGCCCCCCCGCCCCACATCCTGCCCCGCAAAAGGCAGCTTCACCAAAGTGGGGTATTTCC
AGCCTTTGTAGCTTTCACTTCCACATCTACCAAGTGGGCGGAGTGGCCTTCTGTGGACGAATCAGATTCC
TCTCCAGCACCGACTTTAAGAGGCGAGCCGGGGGGTCAGGGTCCCAGATGCACAGGAGGAGAAGCAGGAG
CTGTCGGGAAGATCAGAAGCCAGTCATGGATGACCAGCGCGACCTTATCTCCAACAATGAGCAACTGCCC
```

GenBank

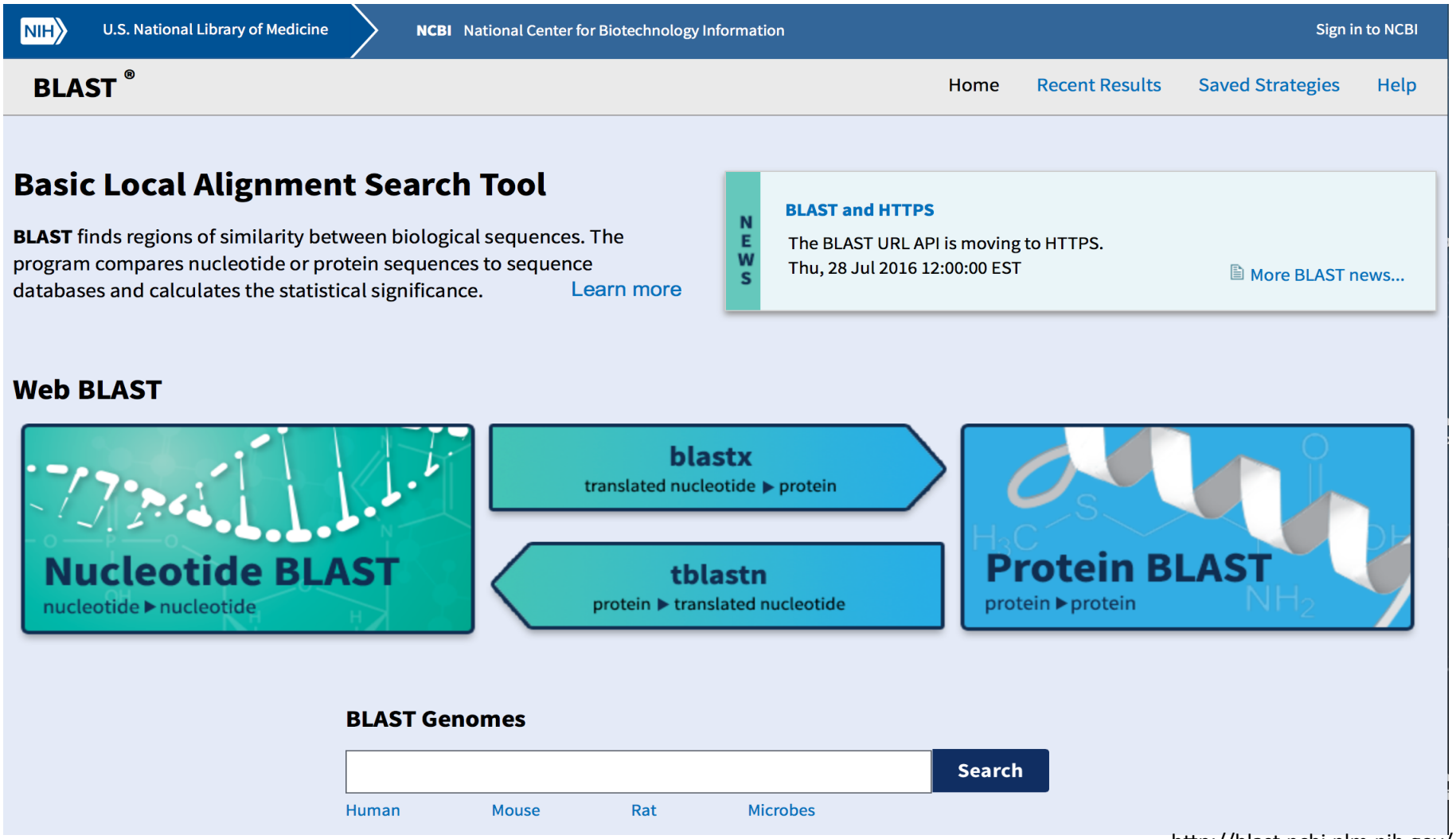
```
  1 ctgcctgggg agccccccg ccccacatcc tgccccgcaa aaggcagctt caccaaagtg
 61 gggtatttcc agcctttgta gctttcactt ccacatctac caagtgggcg gagtggcctt
121 ctgtggacga atcagattcc tctccagcac cgactttaag aggcgagccg gggggtcagg
181 gtcccagatg cacaggagga gaagcaggag ctgtcgggaa gatcagaagc cagtcatgga
```

Accession

```
ACCESSION   NG_029730 REGION: 5001..16300
VERSION     NG_029730.1  GI:343488507
```

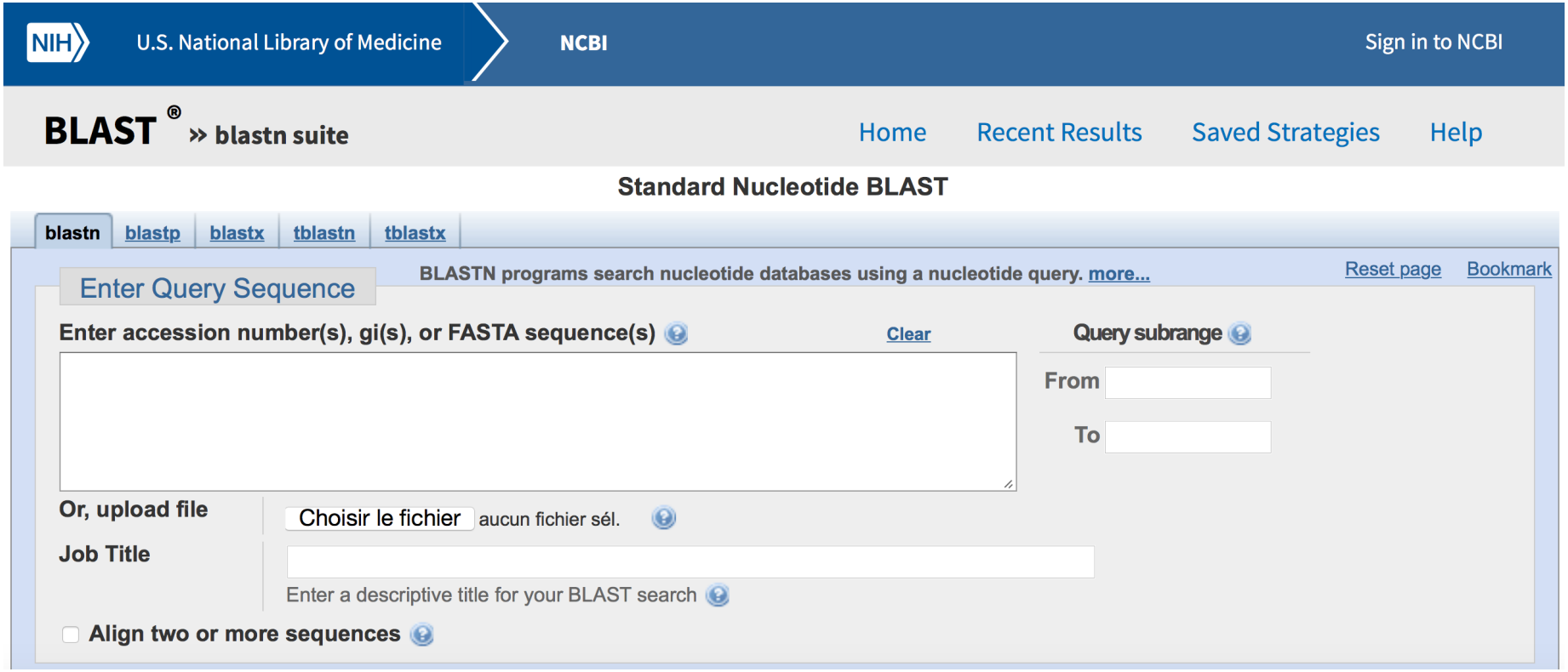
```
>gi|343488507
```

File formats



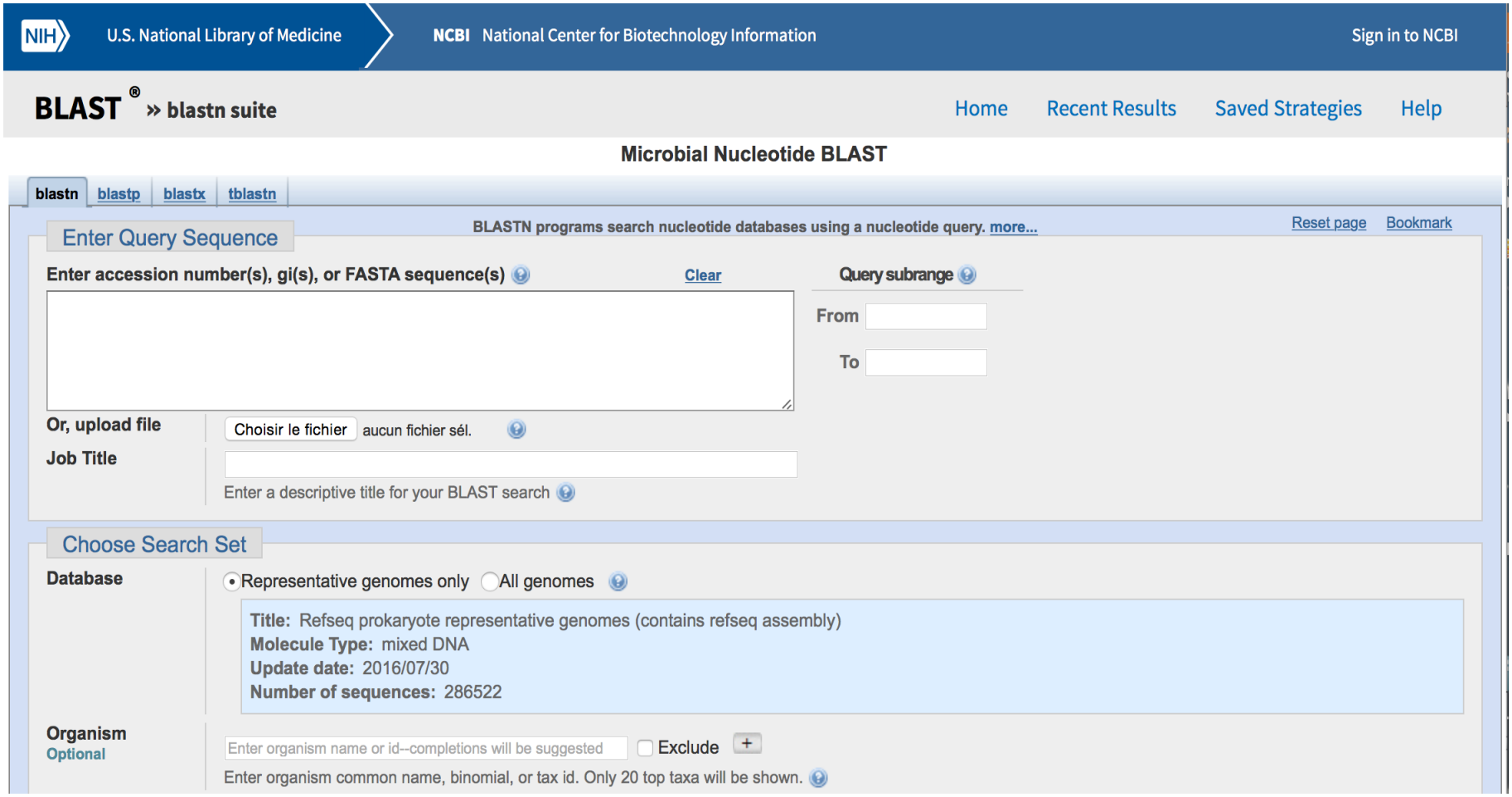
http://blast.ncbi.nlm.nih.gov/

File formats



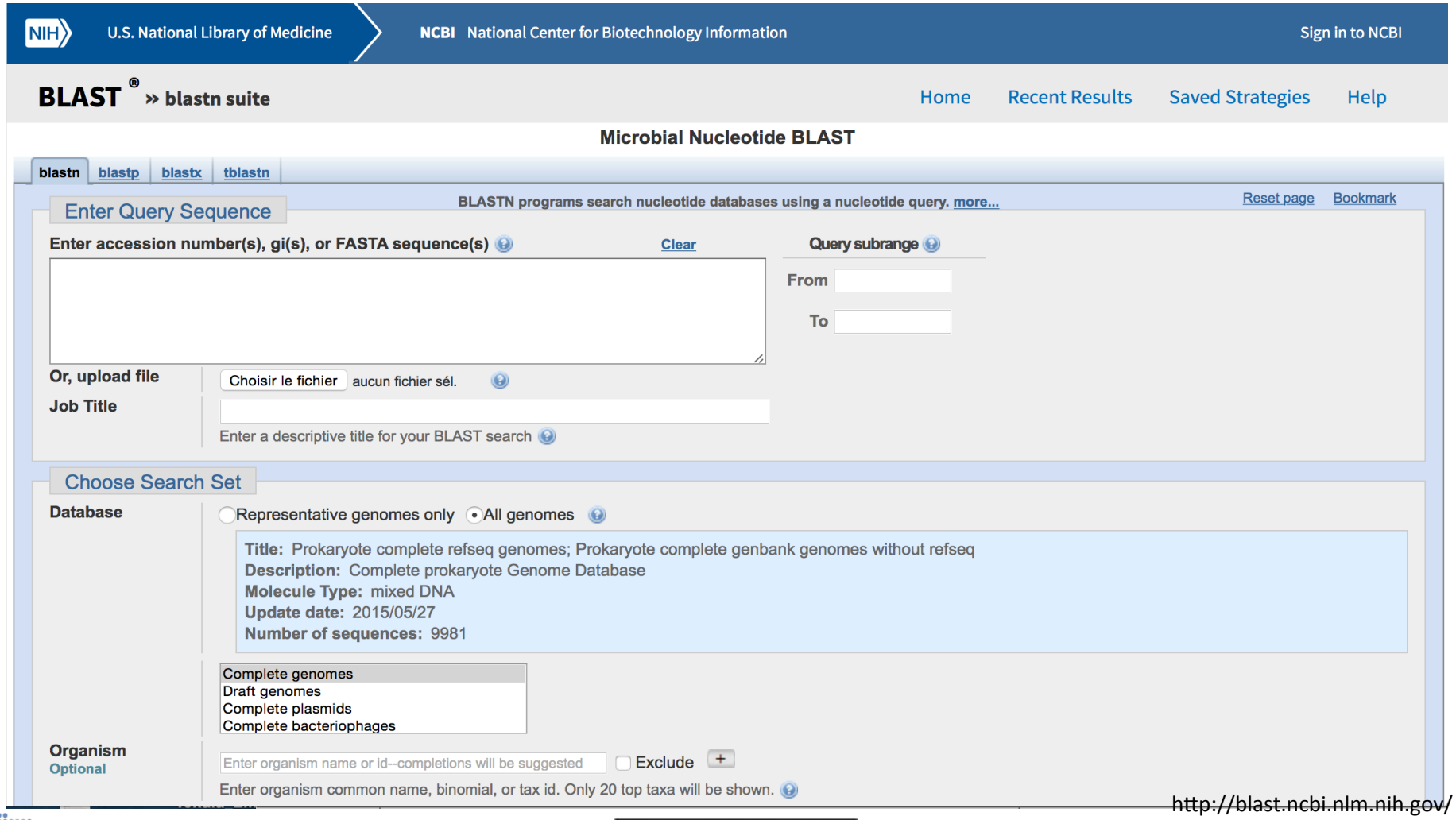
http://blast.ncbi.nlm.nih.gov/

File formats



http://blast.ncbi.nlm.nih.gov/

File formats



http://blast.ncbi.nlm.nih.gov/

H3ABioNet

معهد باستور تونس
Institut Pasteur de Tunis

Part 4

Comparative Genomics: The methods

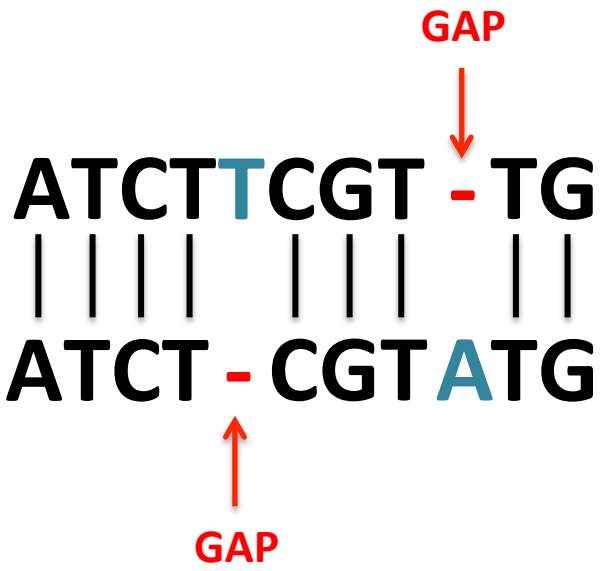
Sequence Alignment for DNA

- DNA sequence alignment consists of aligning 2 DNA sequences in order to identify regions showing sequence similarity

- This highlights regions showing relationship in terms of:
  - Evolution
  - Structure
  - Function

Sequence Alignment for DNA

- For simple sequences:
  Compare **ATCTTCGTTG** and **ATCTCGTATG**



Sequence Alignment for DNA

- For sequences that are not as simple :
  - long sequences
  - more complex sequences (divergent…)
  - large number of sequences
  - …

→ need **Algorithms !**

Sequence Alignment for DNA

- 2 key approaches

- **Global Alignment**

→ Optimizes the alignment to span the full length of sequences that are aligned.

```
ATCATTCGTTGACTGTG
A- - -TT-G - TGAC- - TG
```

- **Local Alignment**

→ Optimizes the alignment to take into account regions of the highest similarity between divergent sequences.

```
ATCATTCGTTGACTGTG
- - - ATT - G-TGACTG - -
```

Sequence Alignment for DNA

- Algorithms efficiency and choice depends on the number of sequences to compare

- **Pairwise Alignment**
  Sequence alignment of **2** sequences
  → Output: function, structure, evolutionary relationship

- **Multiple sequence Alignment**
  Sequence alignment of **3 or more** sequences (same length)
  → Output: homology, evolutionary relationship

Sequence Alignment for DNA

Pairwise Alignment

- **A Pairwise Alignment is** an optimized local or global alignment of **2 sequences.**

- **3 methods:**
  - Dot-matrix
  - Dynamic programming
  - Word-based

NB: Efficiency can be reduced in low complexity regions (repetitive sequences…)
Can be evaluated by the MUM (Maximum Unique Match)
→ Long MUM sequences = more related sequences

Sequence Alignment for DNA

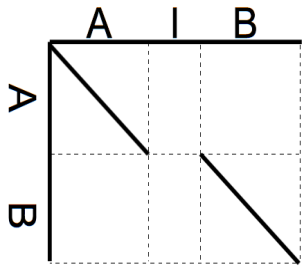
Pairwise Alignment

- **Dot-matrix method**

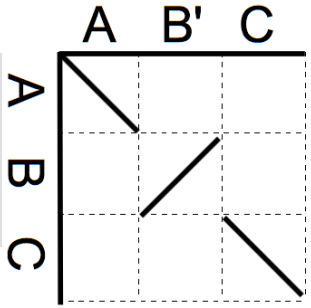
  - 2 sequences (A and B) are aligned using a 2-dimensional matrix
  - Identity is shown with a dot
  - Diagonal shows high similarity

  → Dot plot of the sequence R against sequence Q

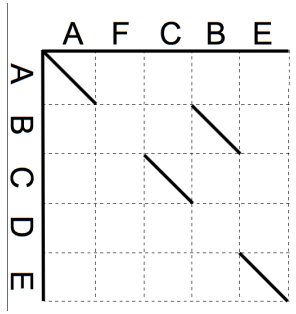
INSERTION INTO QUERY
R: AB
Q: AIB



INVERTION INTO QUERY
R: ABC
Q: AB'C



REARRANGEMENT INTO QUERY
R: ABCDE
Q: AFCBE



Adapted from http://mummer.sourceforge.net/manual/AlignmentTypes.pdf

Sequence Alignment for DNA

Pairwise Alignment

- **Dynamic programming can**

  - Use a scoring matrix
  - Assign a match score (+), a mismatch score (-), and a gap penalty (-).
  - Use two different gap penalties for opening a gap and for extending a gap (gap opening >>> gap extension)
    → generally results in less gaps in an alignment and gaps are grouped together = more biological relevance.

- **Different algorithms for Global and Local Alignments**

https://en.wikipedia.org/wiki/Sequence_alignment#cite_note-mount-1

Sequence Alignment for DNA

Pairwise Alignment

- **Dynamic programming**

- **Global Alignment**

→ Needleman–Wunsch algorithm.

- **Local Alignment**

→ Smith–Waterman algorithm.

https://en.wikipedia.org/wiki/Sequence_alignment#cite_note-mount-1

Sequence Alignment for DNA

Pairwise Alignment

- **Global Alignment**
  → Needleman–Wunsch algorithm.

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

https://en.wikipedia.org/wiki/Needleman–Wunsch_algorithm

Sequence Alignment for DNA

Pairwise Alignment

- **Local Alignment**
  → Smith-Waterman algorithm.

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 1 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 1 & 3 & 2 & 3 & 2 & 3 & 2 & 2 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 4 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 6 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 8 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 10 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

$$T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow \end{pmatrix}$$

https://en.wikipedia.org/wiki/Smith–Waterman_algorithm

Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**

  - Optimal alignment not garanteed, but efficient and faster than dynamic programming
  - Useful for databases searches
  - « words » are small portions (length k) of the query sequence that are used to screen the database

  → Ex: BLAST

Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**

  - **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool)
  - Algorithm to compare a query sequence to a library or database of sequences
  - Allows to estimate identity with a certain confidence threshold
  - Popular in the scientific community (time efficiency…)

Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**

  - **BLAST**

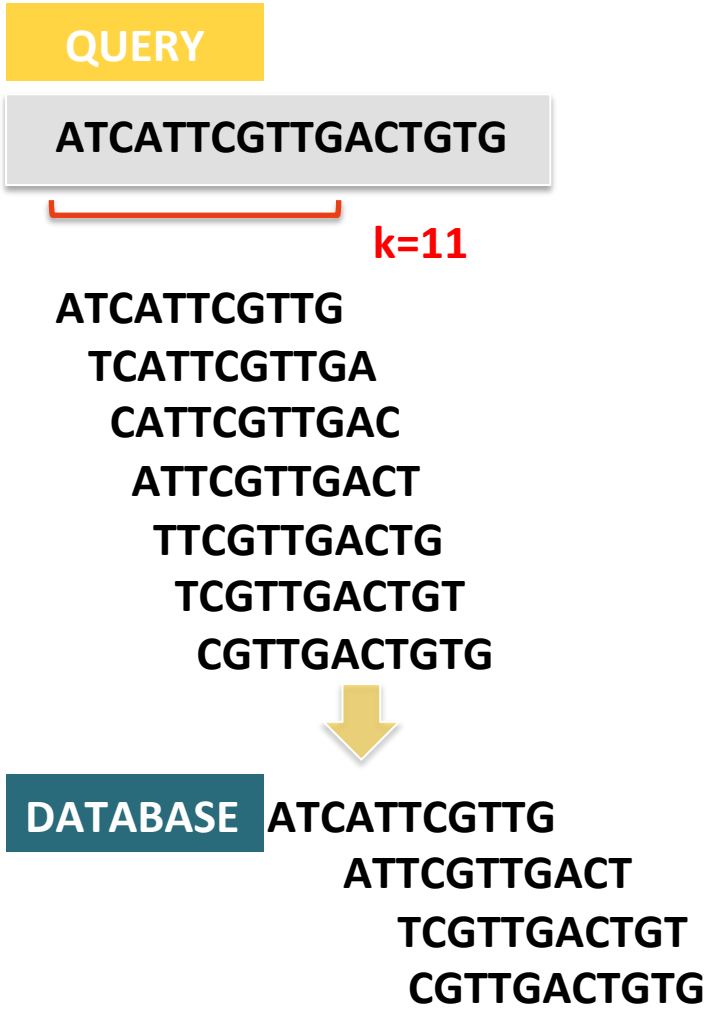
  ▸ The query sequence can be filtered to exclude low-complexity regions

  ▸ **Seeding**: list all possible words of length (DNA: default **k=11**)

  ▸ Search database for matching words using a **scoring matrix = calculate the match score**

  ▸ A **threshold score** is evaluated to top-rank the most similar sequences

  ▸ Process repeated for all words of the query



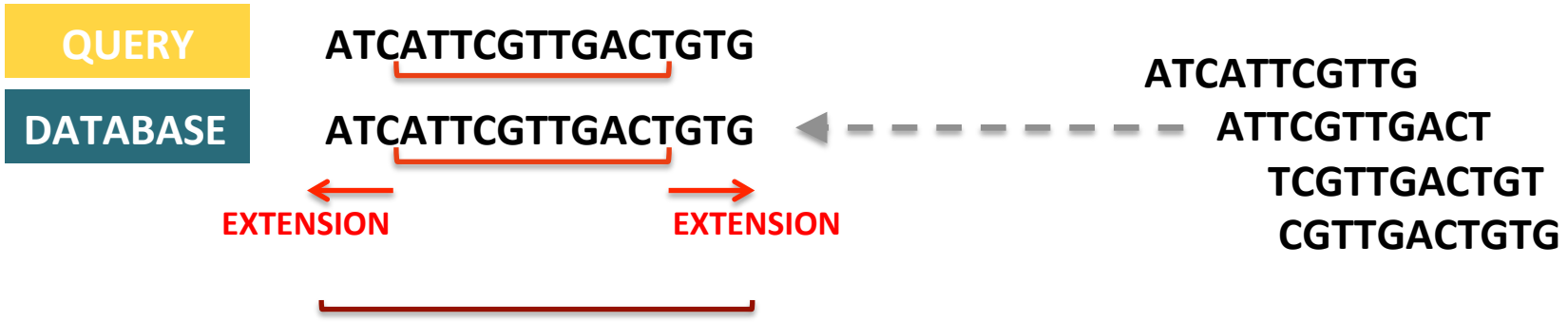
Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**

- **BLAST**

- Best match is followed by an **extension in both direction, with scoring**
- Extension continued only if the alignment is above the threshold
- The contiguous alignment without gaps (now possible) and a higher score is the **HSP** (High Scoring Segment Pair)



Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**

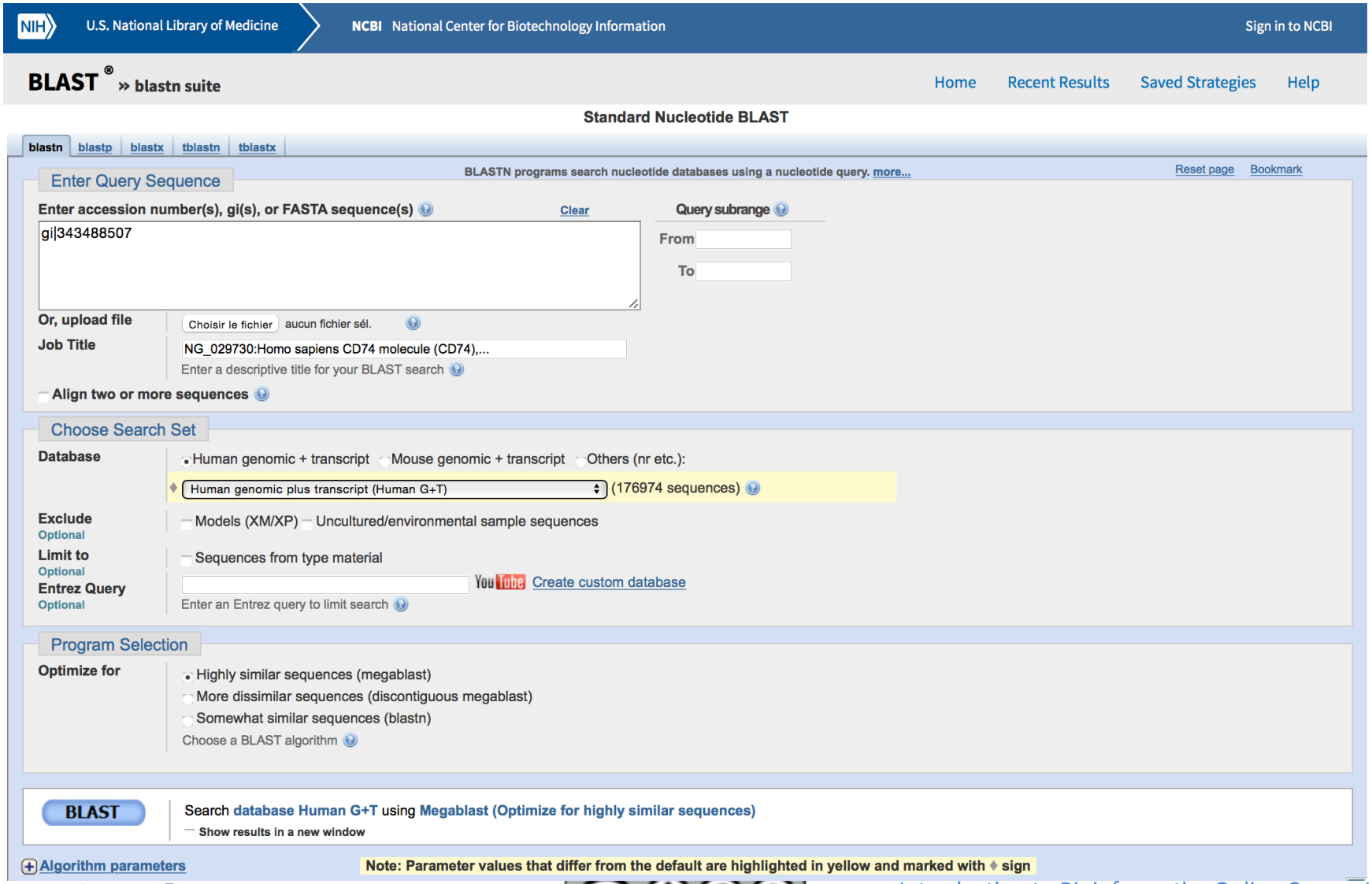
  - **BLAST**

  ▸ A scoring matrix is used to evaluate the quality of the alignment

  ▸ A scoring matrix is a predefined substitution matrix (match = 1, mismatch = 0…)

  ▸ ex: BLOSUM

Sequence Alignment for DNA



http://blast.ncbi.nlm.nih.gov/

Sequence Alignment for DNA

Algorithm parameters

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

Restore default search parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 28

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

Filters and Masking

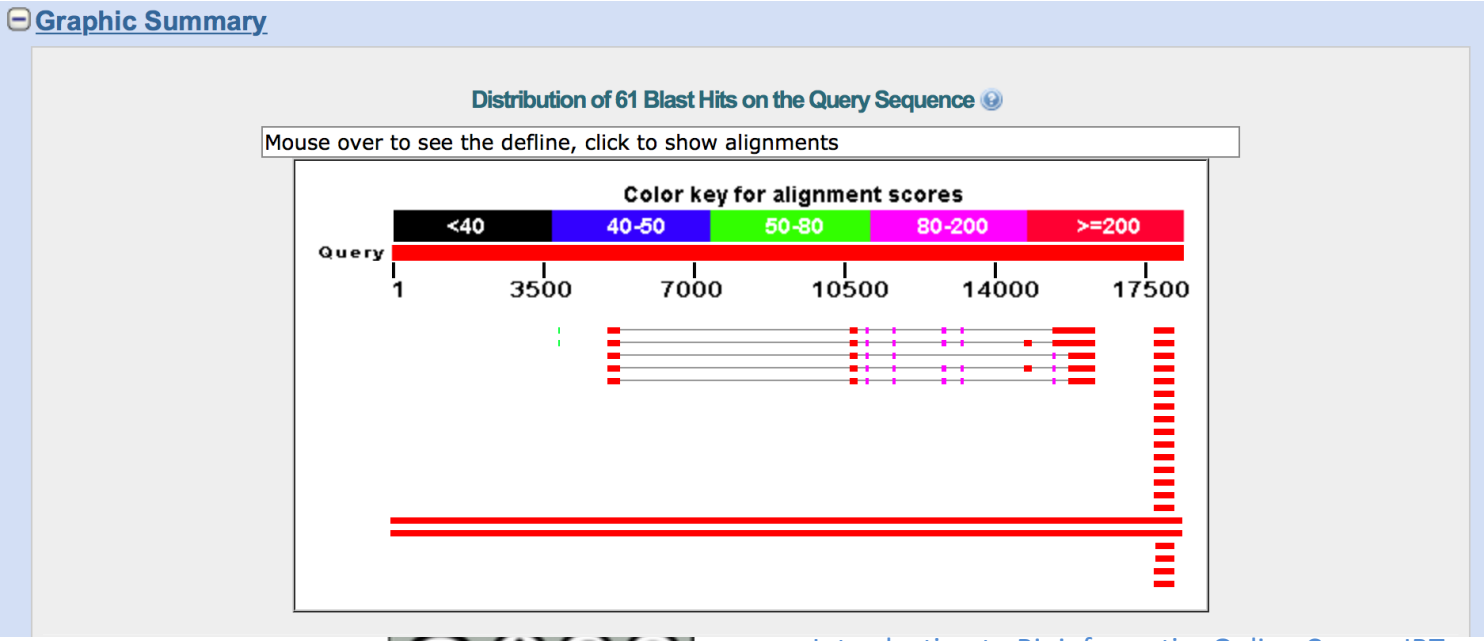
Filter
- ☑ Low complexity regions
- ☐ Species-specific repeats for: Homo sapiens (Human)

Mask
- ☑ Mask for lookup table only
- ☐ Mask lower case letters

Sequence Alignment for DNA

Pairwise Alignment

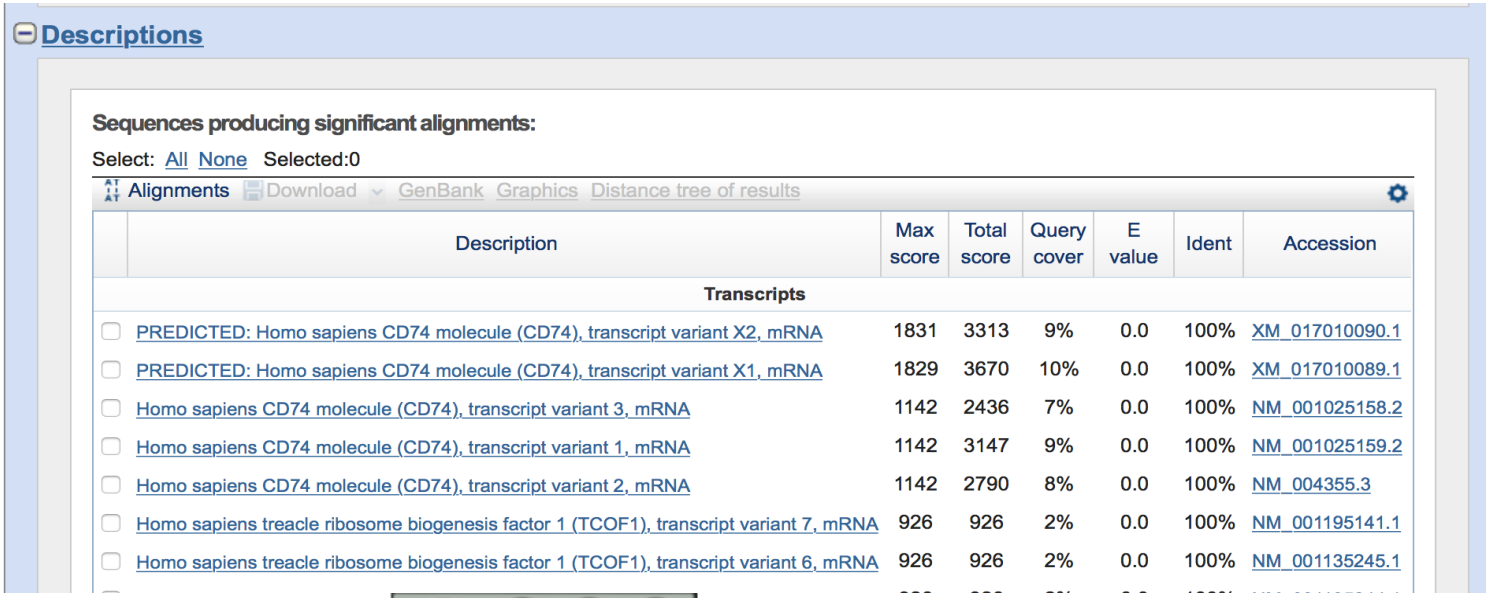
- **Word-based method**
  - **BLAST** output
  - A list of sequences that have the best match to the query



Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**
  - **BLAST** output
  - e-value: probability that the alignment is found by chance (the lower the e-value, the more interesting the match)



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

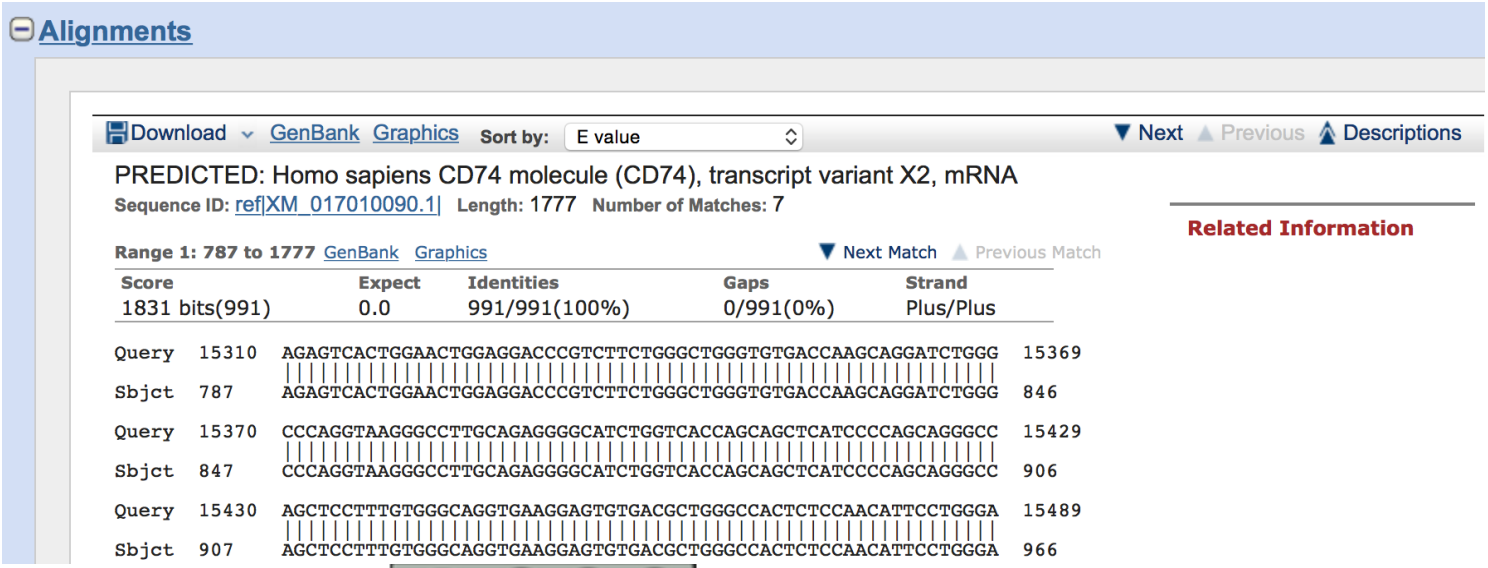
	Description	Max score	Total score	Query cover	E value	Ident	Accession
	Transcripts						
☐	PREDICTED: Homo sapiens CD74 molecule (CD74), transcript variant X2, mRNA	1831	3313	9%	0.0	100%	XM_017010090.1
☐	PREDICTED: Homo sapiens CD74 molecule (CD74), transcript variant X1, mRNA	1829	3670	10%	0.0	100%	XM_017010089.1
☐	Homo sapiens CD74 molecule (CD74), transcript variant 3, mRNA	1142	2436	7%	0.0	100%	NM_001025158.2
☐	Homo sapiens CD74 molecule (CD74), transcript variant 1, mRNA	1142	3147	9%	0.0	100%	NM_001025159.2
☐	Homo sapiens CD74 molecule (CD74), transcript variant 2, mRNA	1142	2790	8%	0.0	100%	NM_004355.3
☐	Homo sapiens treacle ribosome biogenesis factor 1 (TCOF1), transcript variant 7, mRNA	926	926	2%	0.0	100%	NM_001195141.1
☐	Homo sapiens treacle ribosome biogenesis factor 1 (TCOF1), transcript variant 6, mRNA	926	926	2%	0.0	100%	NM_001135245.1

http://blast.ncbi.nlm.nih.gov/

Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**
  - **BLAST** output
  - Alignment details : sequences (query and database) aligned with % identity…)



http://blast.ncbi.nlm.nih.gov/

Sequence Alignment for DNA

Pairwise Alignment

- **Word-based method**

  - BLAST have different variant queries according to the type of query sequence (**Q**) and type of sequence in the database (**R**):

	Q		R
BLAST**N**	Nucleic Acid	→	Nucleic Acid
BLAST**X**	Translated Nucleic Acid	→	Protein
TBLAST**X**	Translated Nucleic Acid	→	Translated Nucleic Acid
TBLAST**N**	Protein	→	Translated Nucleic Acid
BLAST**P**	Protein	→	Protein

Sequence Alignment for DNA

Multiple Sequence Alignment

- **Multiple Sequence Alignment** have been developed to handle more than 2 sequences at a time.
- Align all queried sequences to form a query group.
- Allows to identify **conserved sequences** portions among a group of queried sequences that are:
  - known to be evolutionarily related
  - of unknown/supposed evolutionar relationship → this multiple alignment helps to establish their evolutionar relationships (phylogenetic trees)

Sequence Alignment for DNA

Multiple Sequence Alignment

- **Different methods:**

  - Dynamic programming
  - Progressive method
  - Iterative method : HMMs (Hidden Markov Models)
  - …

→ **Evaluation of Conservation across sequences**

Sequence Alignment for DNA

Multiple Sequence Alignment

- **Dynamic programming**

  - Optimized for 2 sequences, so computationally expensive here
  - Extends the sequence matrix from 2 sequences to the number of sequences in the query : alignment between pairs of sequences

  → MSA

Sequence Alignment for DNA

Multiple Sequence Alignment

- **Progressive method**

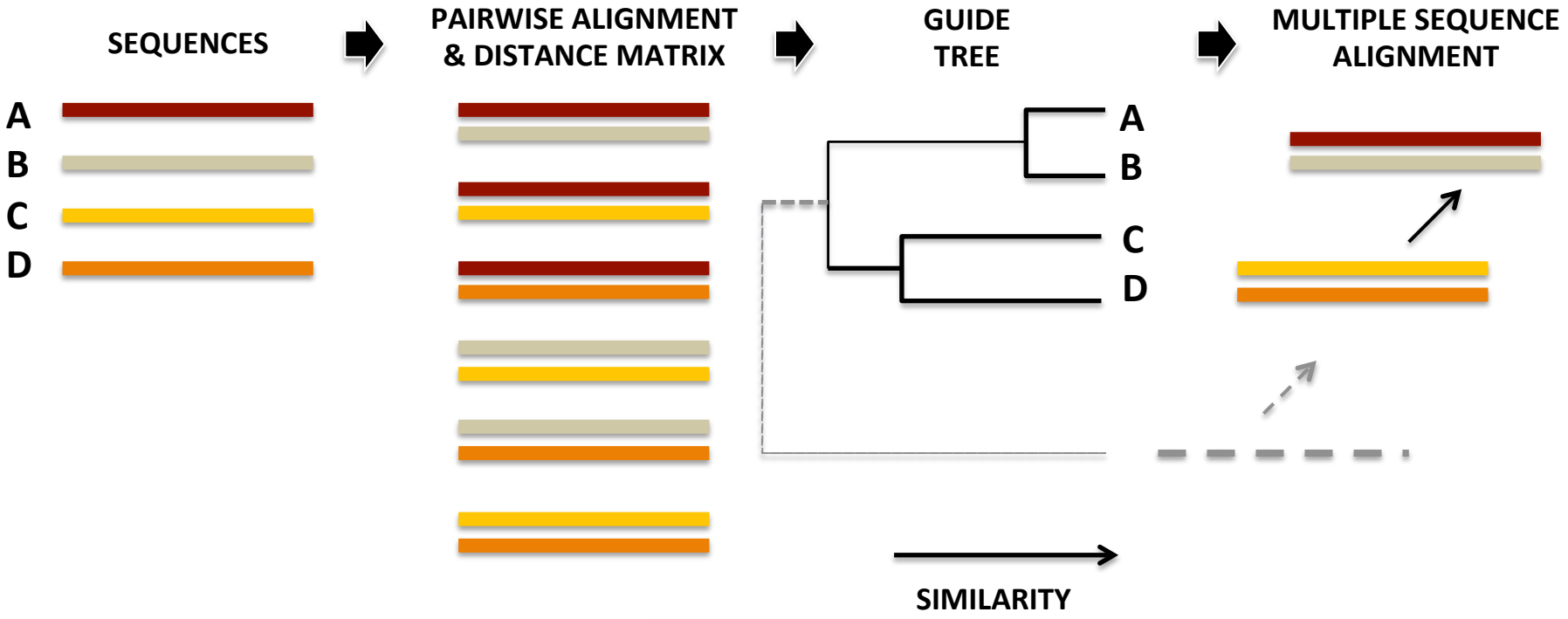
  - Aligns sequences to identify the most similar ones
  - Then progressively adds all other related sequences of the group

  → Clustal (clustal-Omega: medium-large alignments)
  → T-Coffee (small alignments)

Sequence Alignment for DNA

Multiple Sequence Alignment

- **Progressive method**
  Clustal

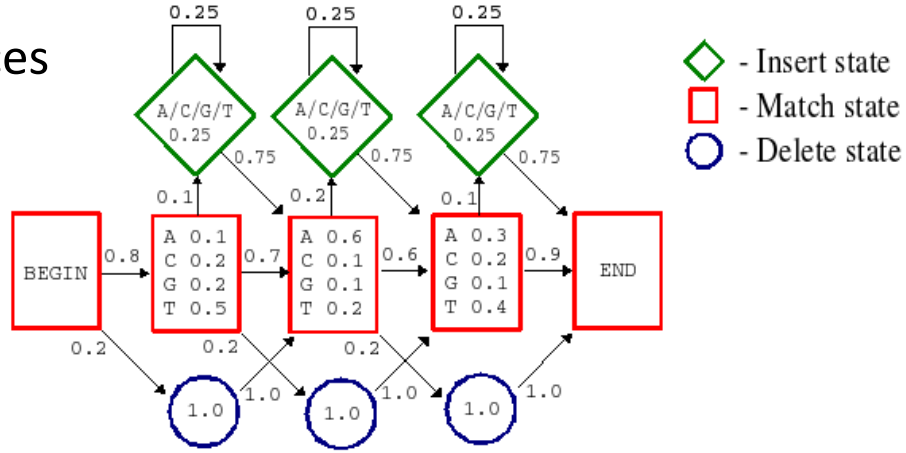


Sequence Alignment for DNA

Multiple Sequence Alignment

- **HMMs**

  - Determine Probability scores for multiple sequence alignments
  - The aligned sequences serves as a group, no need to previously order the sequences
  - Can build an HMM profile
  - Improved for more distant sequences



Sequence Alignment for DNA

Multiple Sequence Alignment

- **Conservation scores**

  - Based on the fact that the highest conservation is maintained through evolution for the most important functions (promoters, essential enzymes, exons…)

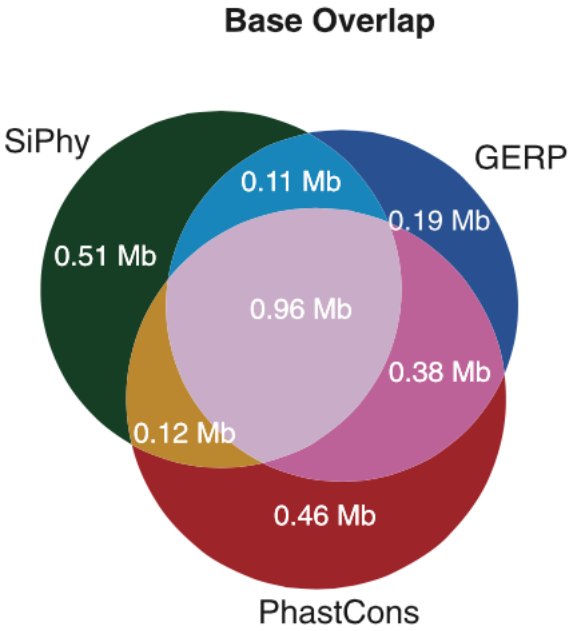
  - Regulatory regions might be generally evolving « faster »

  - Multiple alignments → identify what **elements reject substitutions** (**substitutions occur in neutral DNA**, do not occur if an **element is functionally constrained**)

Sequence Alignment for DNA

Multiple Sequence Alignment

- **Conservation scores**



- **PhyloP** (Phylogenetic p-values)

Measures Base Conservation from non-coding regions

- **PhastCons** (part of PHAST: PHylogenetic Analysis with Space/Time models)

Measures Base Conservation based on HMM model

- **GERP** (Genomic Evolutionary Rate Profiling)

Measures Base Conservation to estimate the neutral evolution rate in genomes

- **SiPhy** (SIte-specific PHYlogenetic analysis)

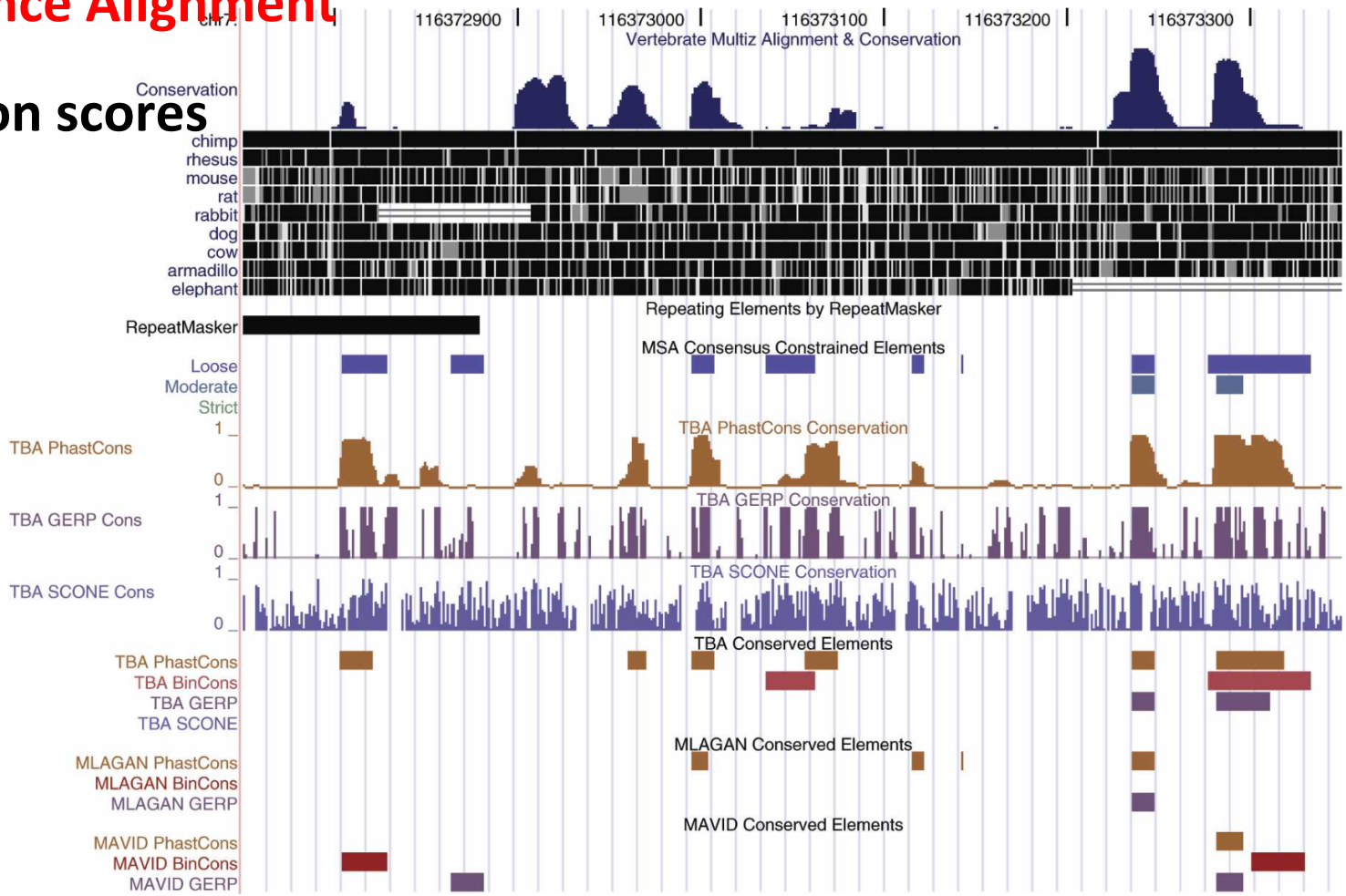
Models the pattern of substitution (biased nucleotide substitution, HMM)

(Gerber et al., 2009)

Sequence Alignment for DNA

Multiple Sequence Alignment

- Conservation scores



(Pheasant and Mattick, 2007)

Sequence Alignment for DNA

Whole Genome Alignment

- **MUMmer Ultra-fast alignment of large-scale DNA**
  (http://mummer.sourceforge.net)
  - Alignment of entire genomes (complete or draft)
  - **Maximal Unique Matcher**: Find the MUMs = subsequences that occur only once in both genomes compared and not extendable anymore
  - Suffix-tree based approach

- **WebACT Artemis Comparison Tool (ACT)**
  (www.webact.org)
  Visualize the alignment of publically available prokaryotic genomes

- **Databases: Ensembl, VISTA**

Part 5

Comparative Genomics: The output

Genomic Structure

- Analyzing a genome structure means to analyze:

  - **At the Genome level**
  - Base composition (%GC, codon bias, nucleotide distribution…)
  - Genome organization (SV events, genomic rearrangements…)
  - Sequence conservation (regulatory elements, repetitive regions…)
  - Synteny (conserved or not)

  - **At the Gene level**
  - Gene order

Genomic Function

- Analyzing functions in a genome means to analyze:

  - **At the non-coding sequence level**
  - Regulatory functions…

  - **At the coding sequence level**

→ Insights into functions
- Compare gene sequences
- Compare protein-coding portions

→ **How ?**

Different algorithms help identify portions of the genome coding for proteins
- *Ab initio* approaches
- Using homology …

Genomic Evolution

- From the Multiple sequence alignments we can infer homology and estimate the evolution distance between sequences/ organisms.

- Analysis could be:
  - Based on entire **Genome** comparisons
  - Based on **Gene** comparisons
  - …

Genomic Evolution

- **Based on entire genome comparison**

  - Phylogenetic relationship between organisms

  - Previous dogma:

"Anything found to be true of E. coli must also be true of elephants"
(Jacques Monod, 1954)

→ Need to be related to its phylogenetic context !!
→ Different outputs expected depending if :
  → Comparing closely related species
  → Comparing evolutionarily distant species

Genomic Evolution: in easy words

COMPARING CLOSELY RELATED SPECIES

COMPARING EVOLUTIONARILY DISTANT SPECIES

http://www.beller.no
http://www.notcot.org/

Comparative Genomics and Genome Evolution in easy words

COMPARING CLOSELY RELATED SPECIES

A: NOT SPICY

B: SPICY

COMPARING EVOLUTIONARILY DISTANT SPECIES

C: NOT SPICY

Adapted from http://recettes-aymen.over-blog.fr/
http://quebueno.be/content/6

Genomic Evolution

- **Based on entire genome comparison**

Examining the dynamics of closely related genomes helps to build therapeutic strategies for Ebola virus:

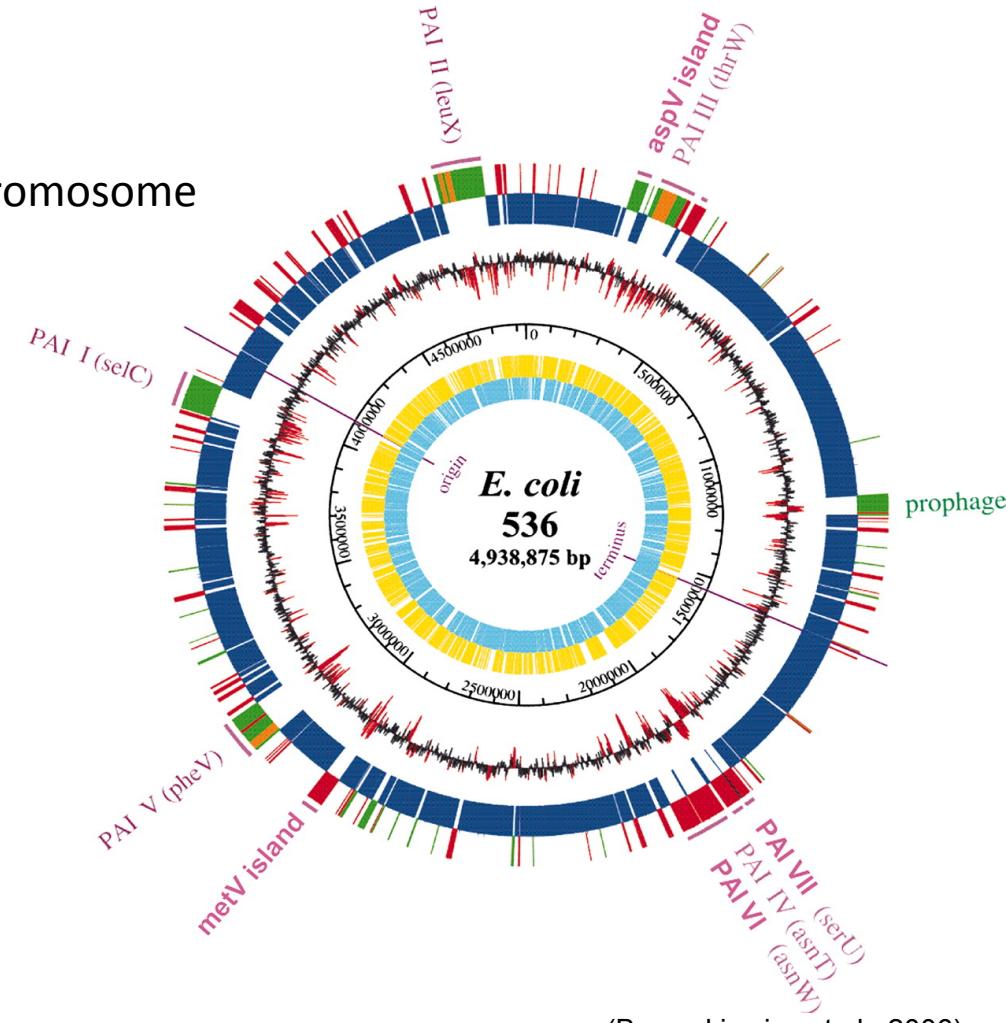
- Ebola virus largest outbreak (2014)
- Comparison of 100 available Ebolavirus (Filiviridae) genomes to each other + to other viral genomes.
- Filoviridae are different from all other viral genomes
- Filovirus genomes : sequence diversity but proteins with similar functions and gene order

→ Ebolavirus genomes very similar but different in intergenic regions and genes of specific function = potential vaccine candidates.

(Jun et al., 2015)

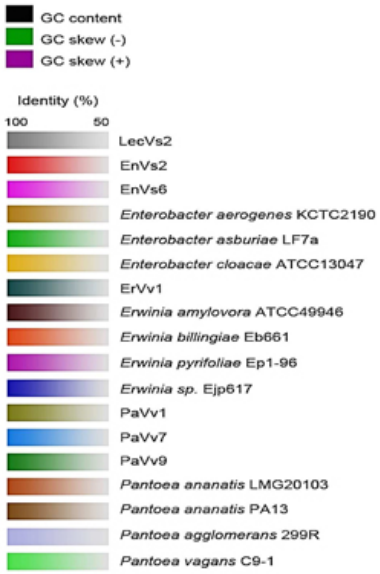
Genomic Structure

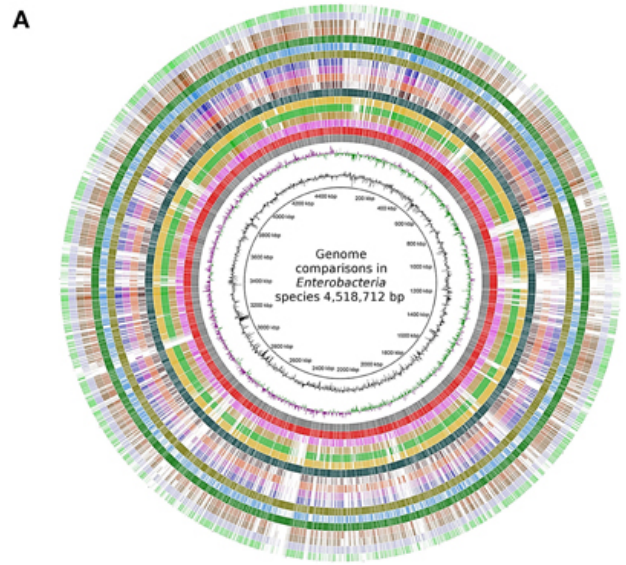
- Genetic map of the UPEC strain 536 chromosome

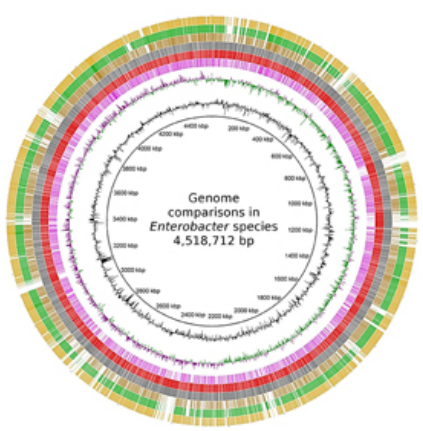


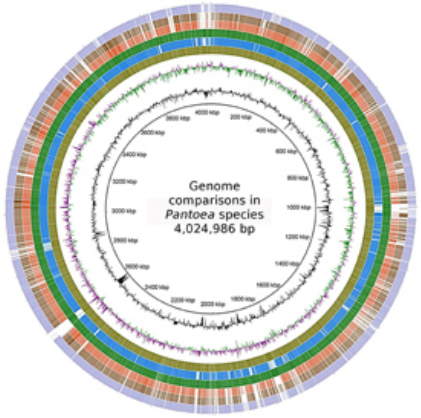
(Brzuszkiewicz et al., 2006)

Genomic Structure









Whole Genome Comparison of
(A) all strains
(B) Enterobacter
(C) Erwinia
(D) Pantoea

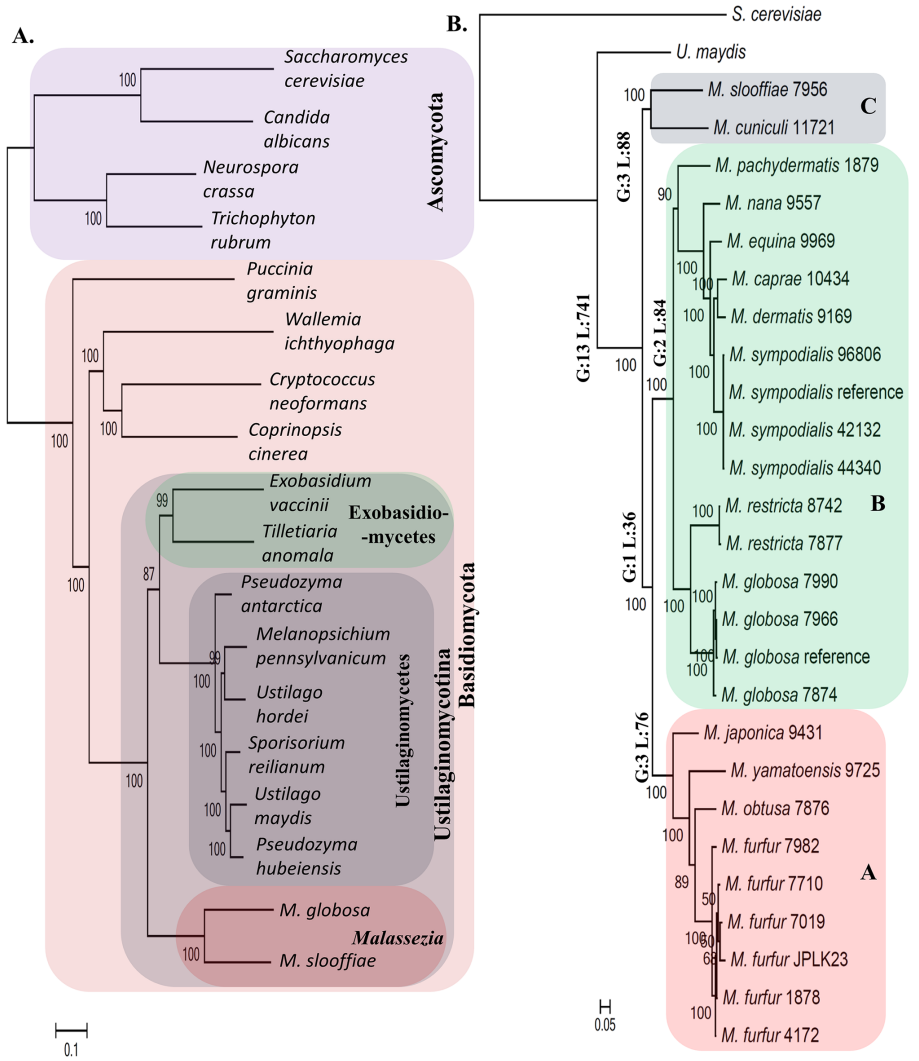
(Lòpez-Fernàndez et al., 2015)

Genomic Evolution

- **Phylogenetic relationships**

« relationship of the Malassezia genus with respect to other fungi with sequenced genomes.»

G: gene family gain; L: gene family loss



(Wu et al., 2015)

https://en.wikipedia.org/wiki/Sequence_alignment#cite_note-mount-1

Genomic Evolution

- **Based on entire genome comparison**

  - **Synteny**

  ▸ Defined as the overall conservation of (gene/blocks) order in chromosomes between different genomes.

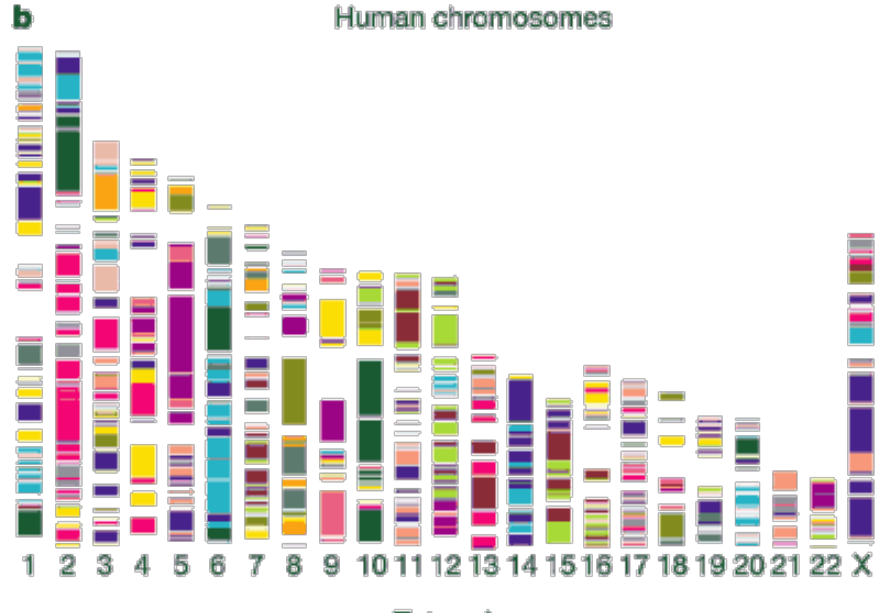
  ▸ Evaluated in whole genomes, blocks could include large portions of genomes.

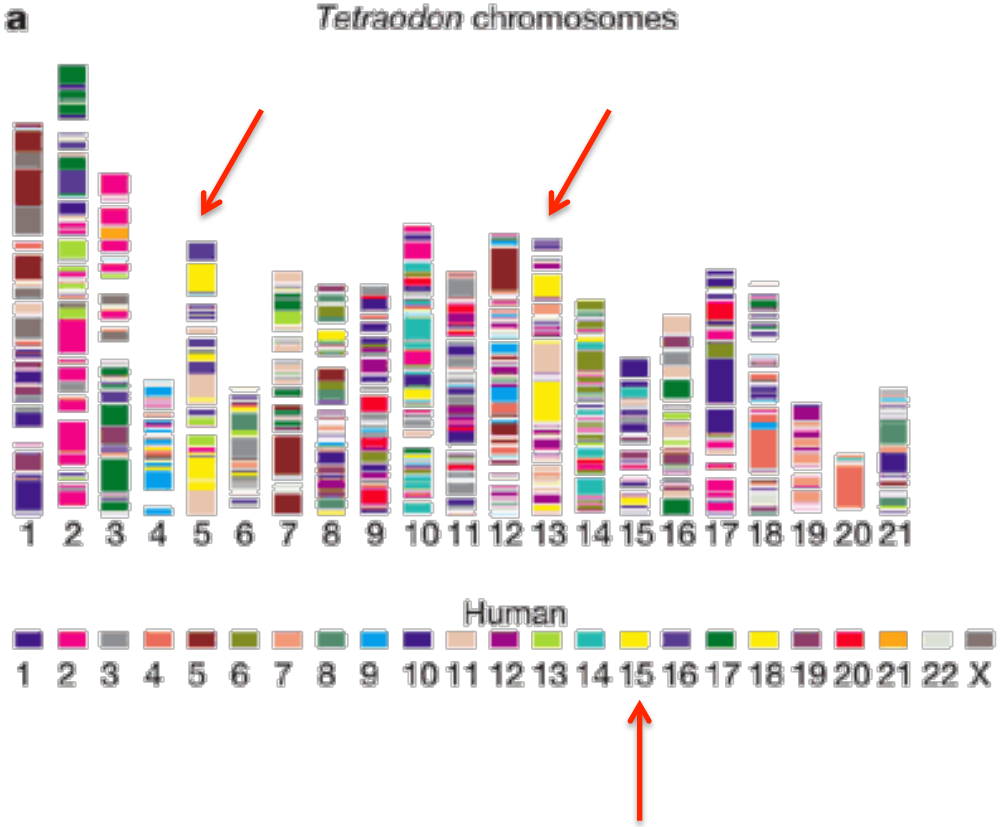
  ▸ Recombination / crossing over affects groups of adjacent genes in a chromosome → **linkage group**.

Genomic Evolution

- Based on entire genome comparison

Synteny

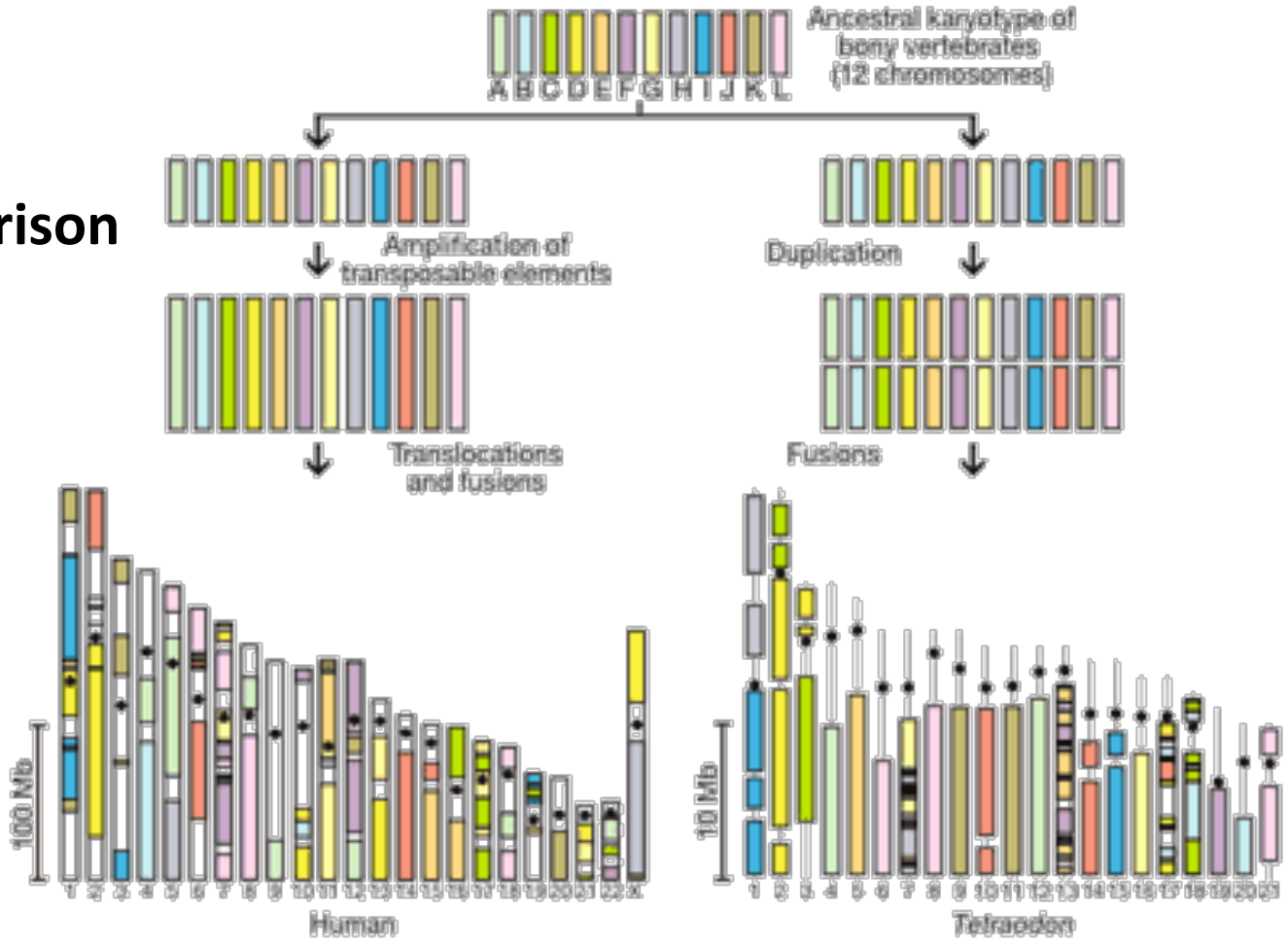




(Jaillon et al., 2004)

Genomic Evolution

- **Based on entire genome comparison**
  Synteny



(Jaillon et al., 2004)

Genomic Evolution

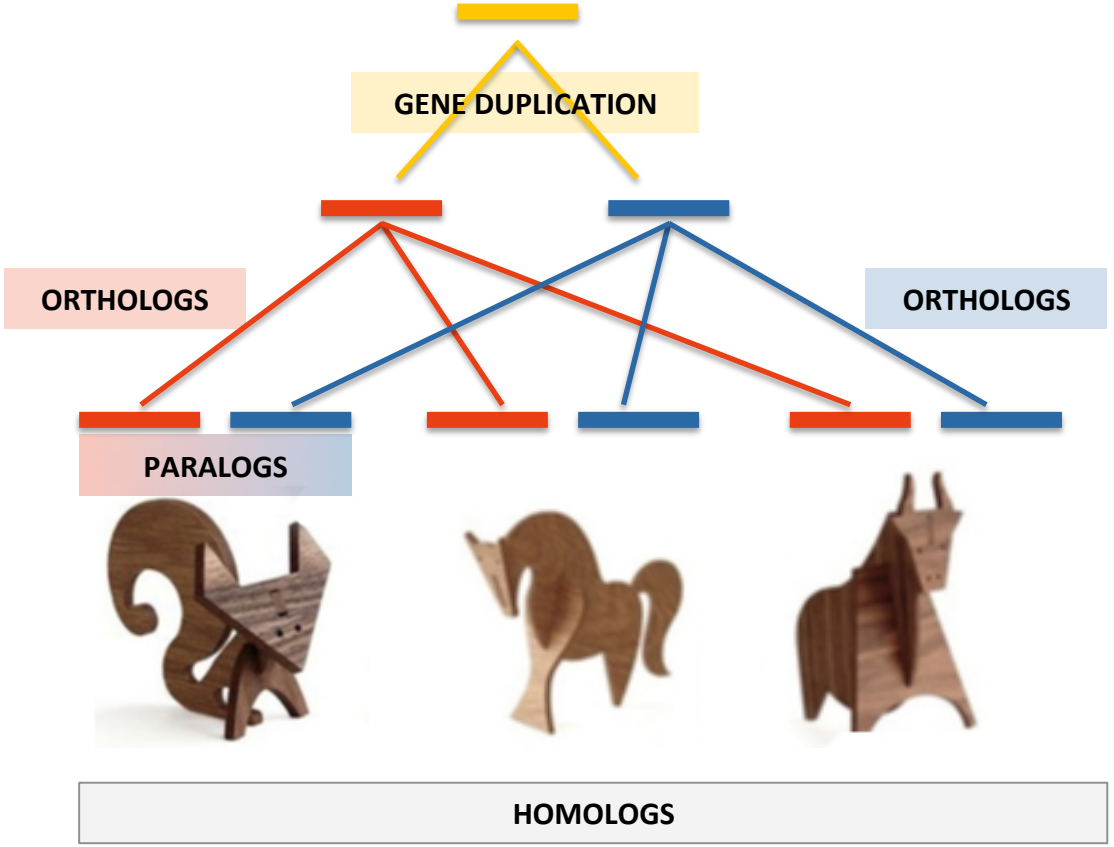
- **Based on individual gene comparison**

Homology
2 genes are homologs if they have a common ancestor

They can be classified in orthologs and paralogs:

→ As a consequence of speciation = **Orthology**

→ As a consequence of duplication = **Paralogy**



Genomic Evolution

- **Based on individual gene comparison**

Orthology

Finding orthologs can be the first step in whole genome alignment

→ BLAST Reciprocal Best Hit (best pairs of orthologs)
→ OrthoMCL (possible predictions for several species)
→ EnsemblCompara (precomputed data): orthology and paralogy predictions based on phylogenies.
→ eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups):OGs of proteins across different taxonomic levels

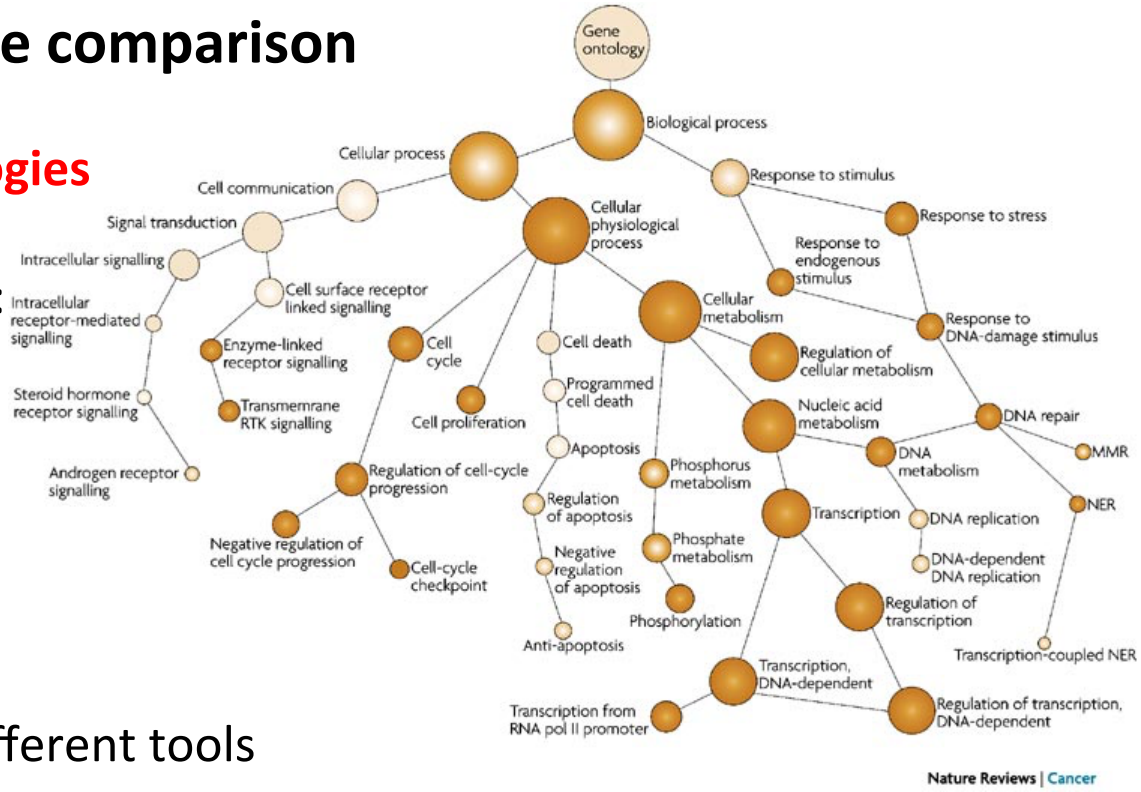
Genomic Evolution

- **Based on individual gene comparison**

Inferring Annotation: Ontologies

Assigning functions to Genes:
GO (Gene Ontology)

→ **Biological Process**
→ **Molecular Function**
→ **Cellular Component**



(Hu et al., 2007)
http://geneontology.org/

Genomic Evolution

- **Functional/Structural Predictions**

▸ **impact of a mutation on the function**

Analysis of the impact of aa substitution

→ Strutural and/or Functional effect of single point mutations SNPs

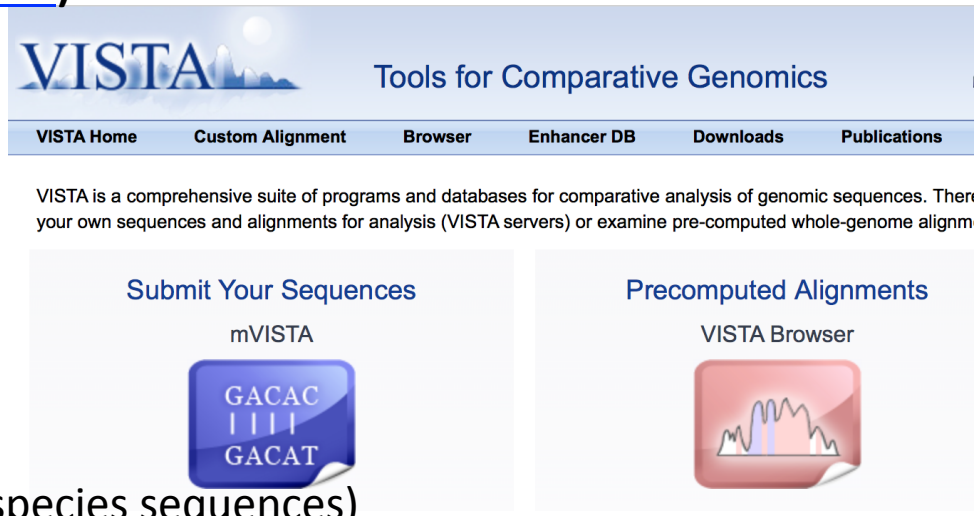
- **PolyPhen-2** (http://genetics.bwh.harvard.edu/pph2/…)
- **SIFT** (http://sift.jcvi.org)
- **VEP** (http://www.ensembl.org/Homo_sapiens/Tools/VEP)

▸ **impact of the mutation on the Struture**

Impact on gene coding portions (gain/loss) or non-coding portions

Browsers

- **VISTA**
  (http://genome.lbl.gov/vista/index.shtml)
  Collection of resources for comparative genomics



- VISTA browers can be used to analyze pre-computed alignments or user generated or queried sequences

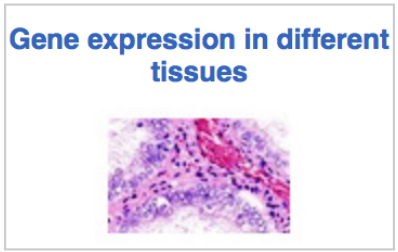
- VISTA servers
  - **mVISTA** (query sequences vs multi-species sequences)
  - **rVISTA** (identification of regulatory TF binding sites)
  - **gVISTA** (query sequences vs whole-genome assemblies)
  - **wgVISTA** (alignment of 10Mb sequences (finished/draft): microbes…) …

- **VISTA** tools: Rviewer region viewer to compare genomic intervals

Browsers

- **Ensembl Browser** (http://www.ensembl.org)

▸ Comparative analyses at the genome and gene levels
▸ Genome sequences compared using pairwise and multiple whole-genome alignments
▸ These alignments help to determine
- Synteny
- Sequence conservation scores
- Gene homology relationships (GeneTrees)



(Herrero et al., 2015)

Comparative genomics
Take-home messages

- **Input / Output**
  - DNA Sequences (genome, gene…)
  - Homology, similarity, evolutionary distance
- **Alignment**
  - Whole genome : MUMmer…
  - Multiple genomes : MGA…
  - Multiple Sequence Alignment : Clustal…
  - Global/Local Sequence Alignment : BLAST…
- **Input / Output files**
  - Fasta/GenBank to alignment or phylogenetic distances