ΒΑΡΒΑΚΕΙΟ ΠΡΟΤΥΠΟ ΛΥΚΕΙΟ

ΟΜΙΛΟΣ ΒΙΟΛΟΓΙΑΣ
«Βιολογία *in vitro & in silico*»

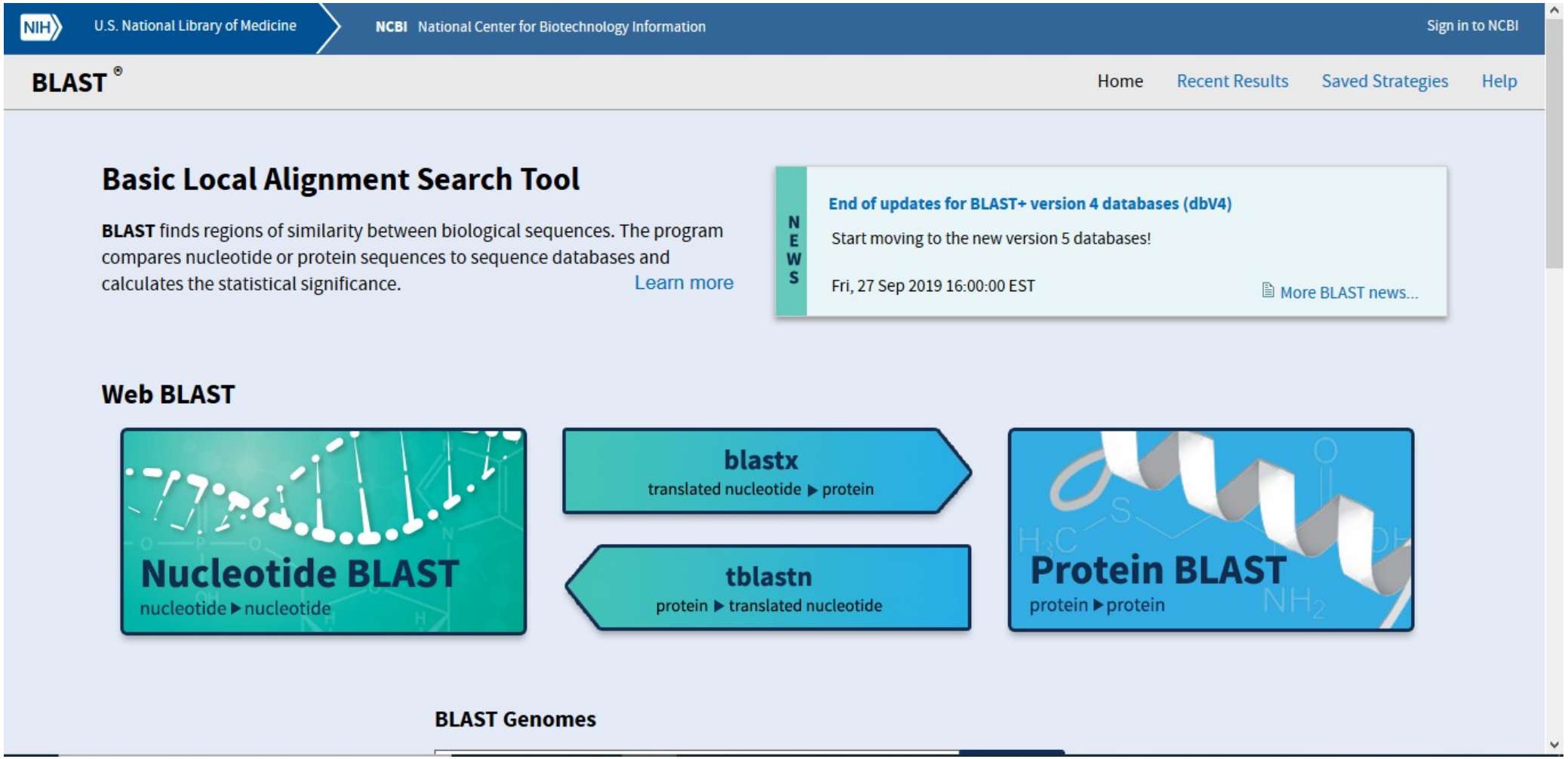
Εισαγωγή στη Βιοπληροφορική

Μαρία Καπασά

Βιολόγος MSc, PhD

BLAST …

Basic Local Alignment Search Tool
Εργαλείο του NCBI για αναζήτηση ομοιότητας ακολουθιών
Ανάλυση σε βάσεις δεδομένων για DNA & πρωτεΐνες
80,000 περίπου αναζητήσεις καθημερινά



Βασικές έννοιες
Ο Αλγόριθμος BLAST (Basic Local Alignment Search Tool)

Σύγκριση μιας αλληλουχίας έναντι των αλληλουχιών μιας ολόκληρης βάσης δεδομένων.

Ταχύτητα, ακρίβεια και ελεύθερη πρόσβαση διαδικτυακά.

Εύρεση ομοιότητας μια νουκλεοτιδικής ή αμινοξικής αλληλουχίας με κάποια/ες άλλες.

Αναζήτηση ομολόγων, ορθολόγων και παραλόγων.

Ανεύρεση νέων, μη χαρακτηρισμένων αλληλουχιών.

Προσδιορισμός πιθανής λειτουργίας μιας αλληλουχίας.

Εύρεση των δομικών λειτουργικών περιοχών των πρωτεϊνών.

Αναζήτηση μικρών εκφρασμένων αλληλουχιών (ESTs).

Αναζήτηση πολυμορφισμών του ενός νουκλεοτιδίου (SNPs).

Τέσσερα βήματα για μια BLAST αναζήτηση

(1) Εύρεση αλληλουχίας- αίτημα (query)

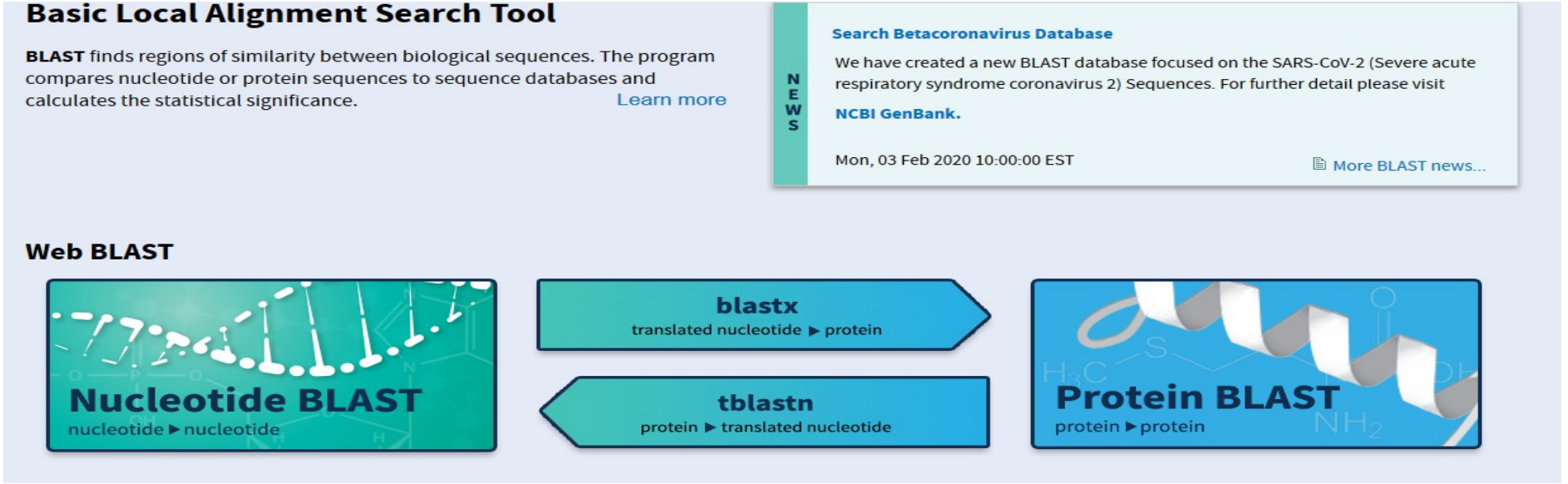
(2) Επιλογή του κατάλληλου τύπου BLAST

(3) Επιλογή της κατάλληλης βάσης δεδομένων

(4) Επιλογή άλλων προαιρετικών παραμέτρων

Και… πατάμε “BLAST”

Βήμα 1: Επιλογή του κατάλληλου τύπου BLAST



Είσοδος προγράμματος			Βάση Δεδομένων
blastn	DNA	1	DNA
blastp	protein	1	protein
blastx	DNA	6	protein
tblastn	protein	6	DNA

Μορφές αλληλουχιών: Παραδείγματα

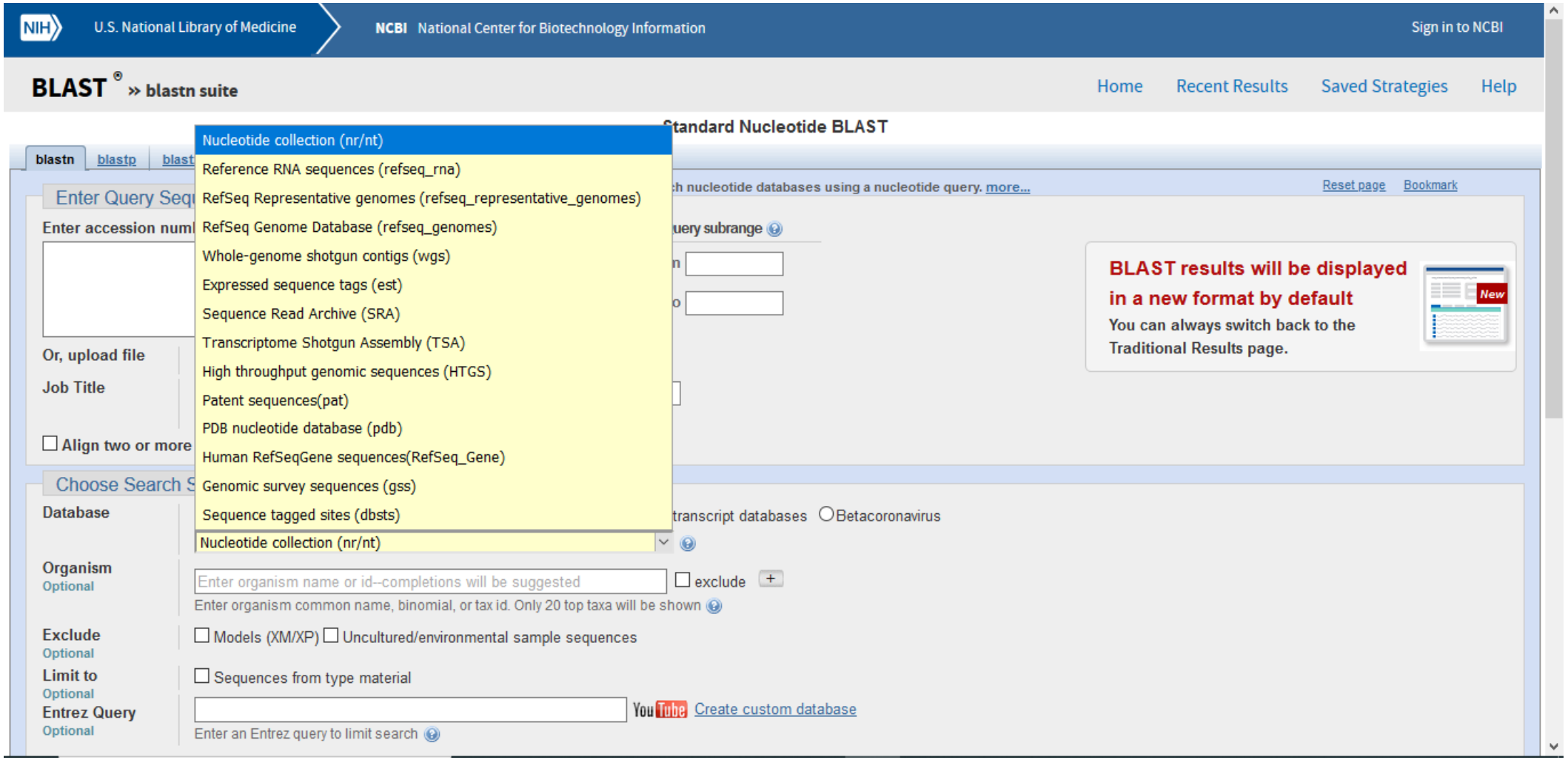
fasta format:

```
>sp|P01588|EPO_HUMAN ERYTHROPOIETIN PRECURSOR - Homo sapiens
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

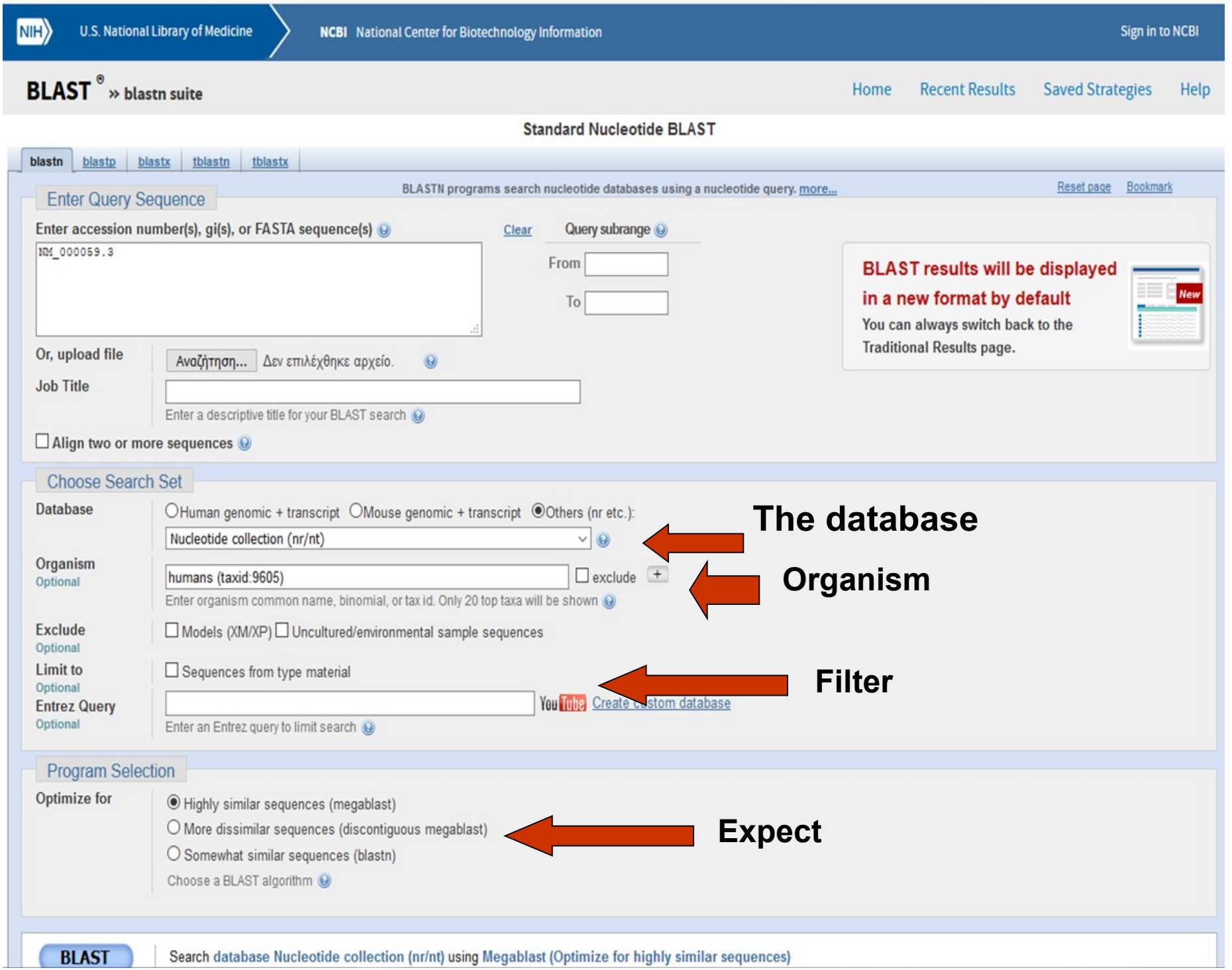
RAW format:

```
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

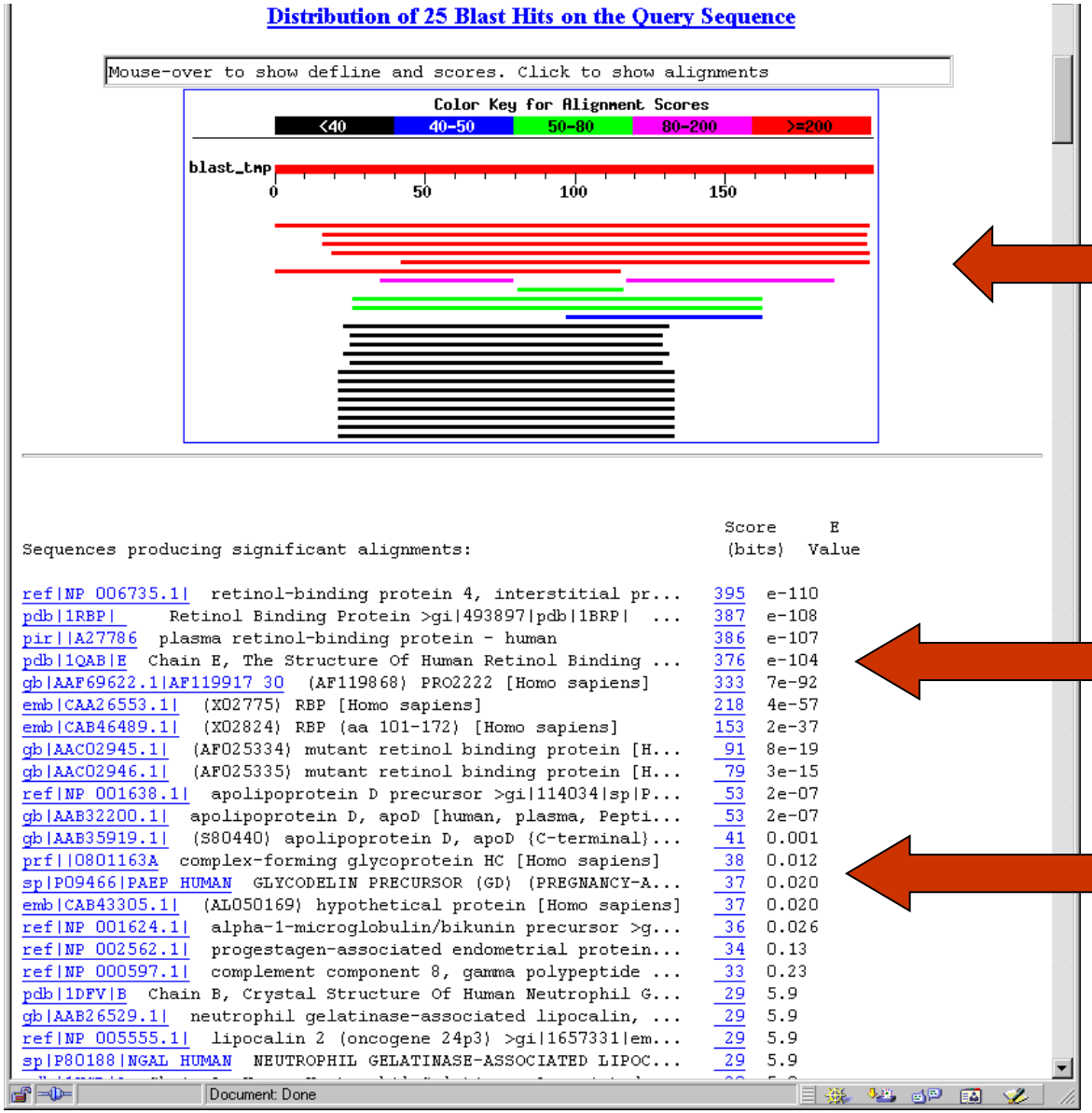
Βήμα 3: Επιλογή της Βάσης Δεδομένων



Βήμα 4: Επιλογή προαιρετικών παραμέτρων



Μορφή παρουσίασης των αποτελεσμάτων BLAST

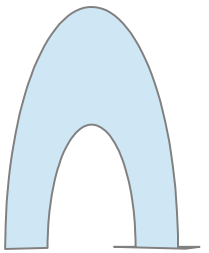


Graphical Overview
Similarity
Length

High scores

Low e values

Cut-off: 0.05?



$e = m * n / 2^{\text{bit-score}}$

m - query sequence length

n - total database length (sum of all sequences)

Bit-score- the required size of a sequence database in which the current match could be found just by chance

Μορφή παρουσίασης των αποτελεσμάτων BLAST

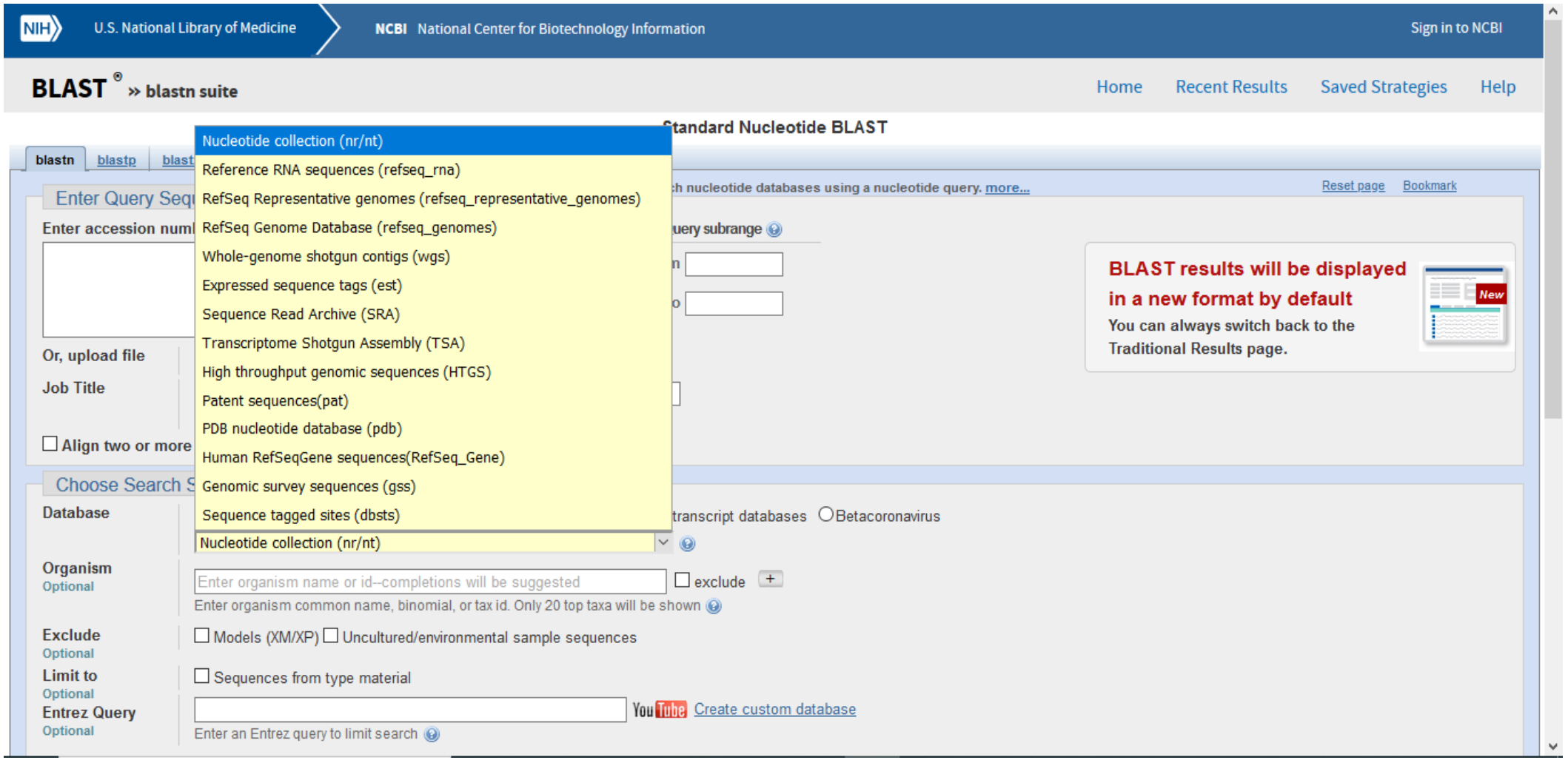
Στοίχιση αλληλουχιών κατά ζεύγη

```
>gi|5803139|ref|NP_006735.1| retinol-binding protein 4, interstitial precursor [Homo sapiens]
 gi|132404|sp|P02753|RETB_HUMAN PLASMA RETINOL-BINDING PROTEIN PRECURSOR (PRBP) (RBP)
 gi|72085|pir||VAHU plasma retinol-binding protein precursor - human
 gi|35897|emb|CAA24959.1| (X00129) precursor RBP [Homo sapiens]
          Length = 199

 Score =  378 bits (971), Expect = e-104
 Identities = 187/199 (93%), Positives = 187/199 (93%)

Query: 1   MKWVXXXXXXXXXXXXERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE 60
           MKWV            ERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE
Sbjct: 1   MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE 60

Query: 61  FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 120
           FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH
Sbjct: 61  FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 120

Query: 121 WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ 180
           WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ
Sbjct: 121 WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ 180

Query: 181 YRLIVHNGYCDGRSERNLL 199
           YRLIVHNGYCDGRSERNLL
Sbjct: 181 YRLIVHNGYCDGRSERNLL 199
```

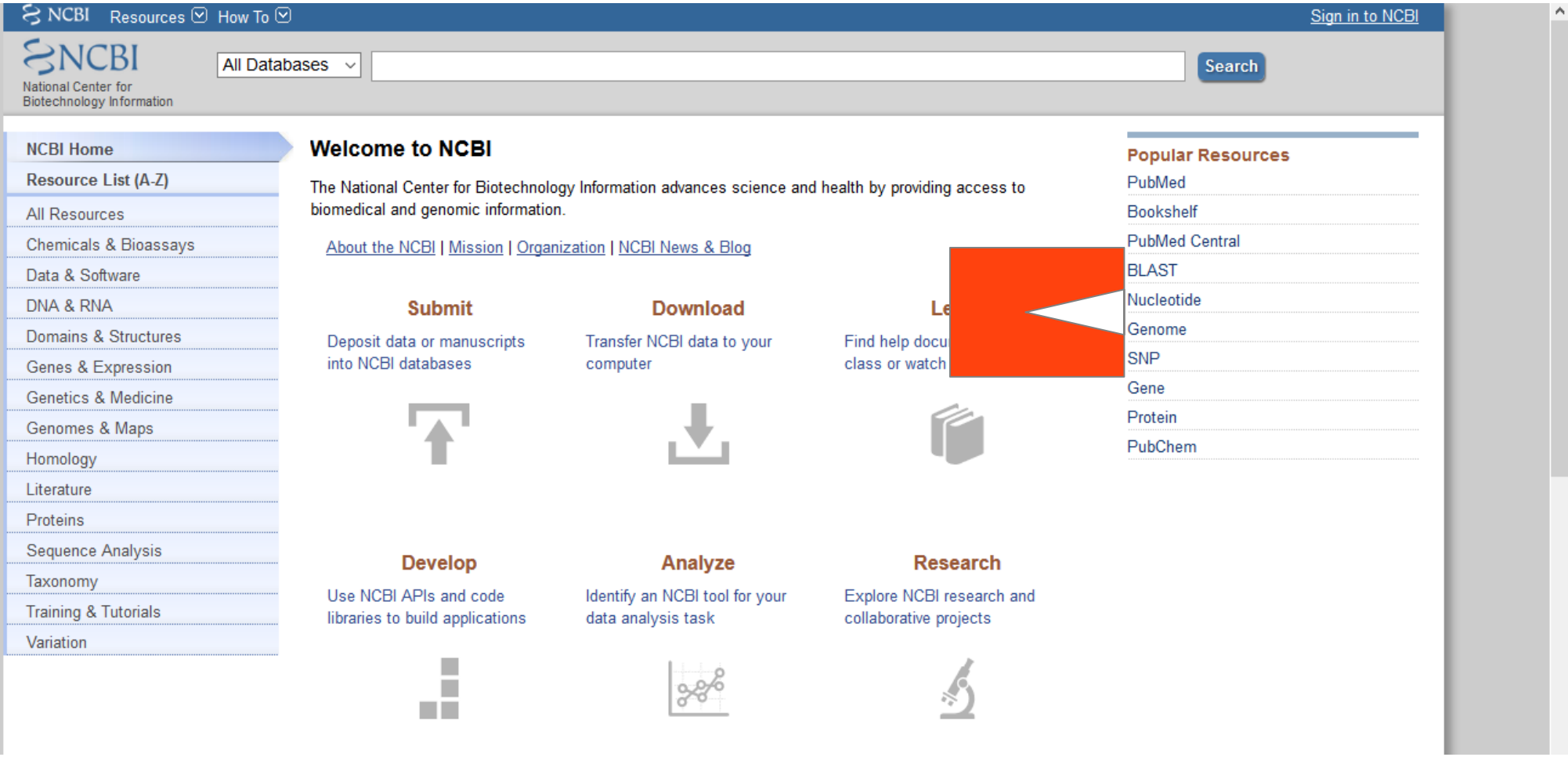
Πολλαπλή στοίχιση αλληλουχιών

```
1_2146    3    WVXXXXXXXXXXXXERDCRVSSFRVKENFDKARFSGTWYAMA-------------KKDPE 49
132407    5    WALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIA-------------KKDPE 51
6978523   33                          VQENFDVKKYLGRWYEI--------------EKIPV 54
6978497   29                        QVQENFNEARIYGKWFNLAVGSTCPWLRRIKNKMSVS 65
266619    35                      SVPLQPGFWTERFQGRWFVVGLAANAV-------QKERQ 66
6981430   30                         VQPNFQQDKFLGRWYSAGLASNSSWF-----REKKE 60
204063    20                         VVKDFDISKFLGFWYEIAFAS----------KMGTP 45
13162312  24                         VVKDFDISKFLGFWYEIAFAS----------KMGTP 49
576102    2                          VVKDFDISKFLGFWYEIAFAS----------KMGTP 27
```

Βήμα 3: Επιλογή της Βάσης Δεδομένων



www.ncbi.nlm.nih.gov



NCBI Resources How To Sign in to NCBI

Nucleotide

Nucleotide BRCA2 HUMAN Search

Create alert Advanced Help

Species
Animals (1,776)
Plants (80)
Fungi (155)
Protists (65)
Bacteria (176)
Viruses (2)
Customize ...

Molecule types
genomic DNA/RNA (2,028)
mRNA (563)
Customize ...

Source databases
INSDC (GenBank) (1,994)
RefSeq (648)
Customize ...

Sequence Type
Nucleotide (2,624)
EST (35)
GSS (1)

Genetic compartments
Mitochondrion (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

GENE Was this helpful?

BRCA2 – BRCA2 DNA repair associated

Homo sapiens (human)

Also known as: BRCC2, BROVCA2, FACD, FAD, FAD1, FANCD, FANCD1, GLM3, PNCA2, XRCC11

GeneID: 675

RefSeq transcripts (1) RefSeq proteins (1) RefSeqGene (1) PubMed (1,766)

Orthologs Genome Browser BLAST Download

[redacted]quences +

Items: 1 to 20 of 2660

<< First < Prev Page 1 of 133 Next > Last >>

1. **Human** breast cancer susceptibility (**BRCA2**) mRNA, complete cds
10,987 bp linear mRNA
Accession: U43746.1 GI: 1161383
Protein PubMed Taxonomy
GenBank FASTA Graphics

2. H.sapiens **brca2** gene exon 2 (and joined coding region)
1,106 bp linear DNA
Accession: X95152.1 GI: 1171546
Protein PubMed Taxonomy
GenBank FASTA Graphics

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(1110)*
unidentified *(384)*
Escherichia coli *(163)*
Mus musculus *(83)*
Rattus norvegicus *(29)*
All other taxa *(891)*
More...

Find related data

Database: Select

Find items

Search details

```
BRCA2[All Fields] AND ("Homo
sapiens"[Organism] OR HUMAN[All
Fields])
```

Search See more...

Recent activity

Turn Off Clear

BRCA2 HUMAN (2660) Nucleotide

BRCA2 (11289) Nucleotide

See more...

NCBI Orthologs How was this calculated?

0 items

SEARCH THE TAXONOMY TREE

Enter taxonomic name

- jawed vertebrates
  - birds
  - alligators and others
  - turtles
  - lizards
  - mammals
  - amphibians
  - coelacanths
  - bony fishes
  - cartilaginous fishes

Genes | Literature

268 genes for: jawed vertebrates (*Gnathostomata*)

Add to cart | Protein alignment | Download

2 selected.

	Species	Gene	Architecture	aa	
☑	*Homo sapiens* human	BRCA2 BRCA2 DNA repair associated		3,418	⌄
☑	*Mus musculus* house mouse	Brca2 breast cancer 2, early onset		3,329	⌄
☐	*Rattus norvegicus* Norway rat	Brca2 BRCA2, DNA repair associated		3,343	⌄
☐	*Canis lupus familiaris* dog	BRCA2 BRCA2 DNA repair associated		3,446	⌄
☐	*Danio rerio* zebrafish	brca2 BRCA2 DNA repair associated		2,874	⌄

NCBI Orthologs How was this calculated?

0 items

Genes | Literature

SEARCH THE TAXONOMY TREE

Enter taxonomic name

- jawed vertebrates
  - birds
  - alligators and others
  - turtles
  - lizards
  - mammals
  - amphibians
  - coelacanths
  - bony fishes
  - cartilaginous fishes

268 genes for: jawed vertebrates (*Gnathostomata*)

Add to cart | Protein alignment | Download

2 selected.

Species		aa
Homo sapiens human		3,418
Mus musculus house mouse	breast cance early onset	3,329

Protein alignment

- one sequence per gene (2)
- all sequences per gene (6)

Align

1 - 3,461 (3,461r shown) Download Tools

Sequence ID	Start	1 500 1000 1500 2000 2500 3000 3461	End	Organism
NP_000050.2	1		3,418	Homo sapiens
NP_033895.2	1		3,329	Mus musculus

PROTEIN: 1 - 3,461 (3,461r shown)

Descriptions Select All Re-align Alignment parameters

Accession	Description	Links
NP_000050.2	breast cancer type 2 susceptibility protein [Homo sapiens]	Related Information
NP_033895.2	breast cancer type 2 susceptibility protein homolog [Mus musculus]	Related Information

Alignments Select All Re-align Mouse over the sequence identifer for sequence title

View Format: Compact Conservation Setting: 2 Bits

```
NP_000050.2  1    MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEELSSEAPPYNSEPAEESEHKNNNYEPNLFKTPQRKPSYNQLASTPII  80
NP_033895.2  1    MPVEYKRRPTFWEIFKARCSTADLGPISLNWFEELSSEAPPYNSEPPEESEYKPHGYEPQLFKTPQRNPPYHQFASTPIM  80

NP_000050.2  81   FKEQGLTLPLYQSPVKELDKFKLDLGRNVPNSRHKSLRTVKTKMDQADDVSCPLLNSCLSESPVVLQCTHVTPQRDKSVV  160
NP_033895.2  81   FKERSQTLPLDQSPFREL-------GKVVASSKHKTHSKKKTKVDPVVDVASPPLKSCLSESPLTLRCTQAVLQREKPVV  153

NP_000050.2  161  CGSLFHTPKFVKGRQTPKHISESLGAEVDPDMSWSSSLATPPTLSSTVLIVRNEEASETVFPHDTTANVKSYFSNHDESL  240
NP_033895.2  154  SGSLFYTPKLKEG-QTPKPISESLGVEVDPDMSWTSSLATPPTLSSTVLIARDEEARSSVTPADSPATLKSCFSNHNESP  232
```

NCBI Resources How To Sign in to NCBI

Nucleotide

Nucleotide BRCA2 HUMAN Search

Create alert Advanced Help

Species
Animals (1,776)
Plants (80)
Fungi (155)
Protists (65)
Bacteria (176)
Viruses (2)
Customize ...

Molecule types
genomic DNA/RNA (2,028)
mRNA (563)
Customize ...

Source databases
INSDC (GenBank) (1,994)
RefSeq (648)
Customize ...

Sequence Type
Nucleotide (2,624)
EST (35)
GSS (1)

Genetic compartments
Mitochondrion (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

GENE Was this helpful?

BRCA2 – BRCA2 DNA repair associated

Homo sapiens (human)

Also known as: BRCC2, BROVCA2, FACD, FAD, FAD1, FANCD, FANCD1, GLM3, PNCA2, XRCC11

GeneID: 675

RefSeq transcripts (1) RefSeq proteins (1) RefSeqGene (1) PubMed (1,766)

Orthologs Genome Browser BLAST Download

ences +

Items: 1 to 20 of 2660

<< First < Prev Page 1 of 133 Next > Last >>

1. **Human** breast cancer susceptibility (**BRCA2**) mRNA, complete cds
10,987 bp linear mRNA
Accession: U43746.1 GI: 1161383
Protein PubMed Taxonomy
GenBank FASTA Graphics

2. H.sapiens **brca2** gene exon 2 (and joined coding region)
1,106 bp linear DNA
Accession: X95152.1 GI: 1171546
Protein PubMed Taxonomy
GenBank FASTA Graphics

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(1110)*
unidentified *(384)*
Escherichia coli *(163)*
Mus musculus *(83)*
Rattus norvegicus *(29)*
All other taxa *(891)*
More...

Find related data

Database: Select

Find items

Search details

```
BRCA2[All Fields] AND ("Homo
sapiens"[Organism] OR HUMAN[All
Fields])
```

Search See more...

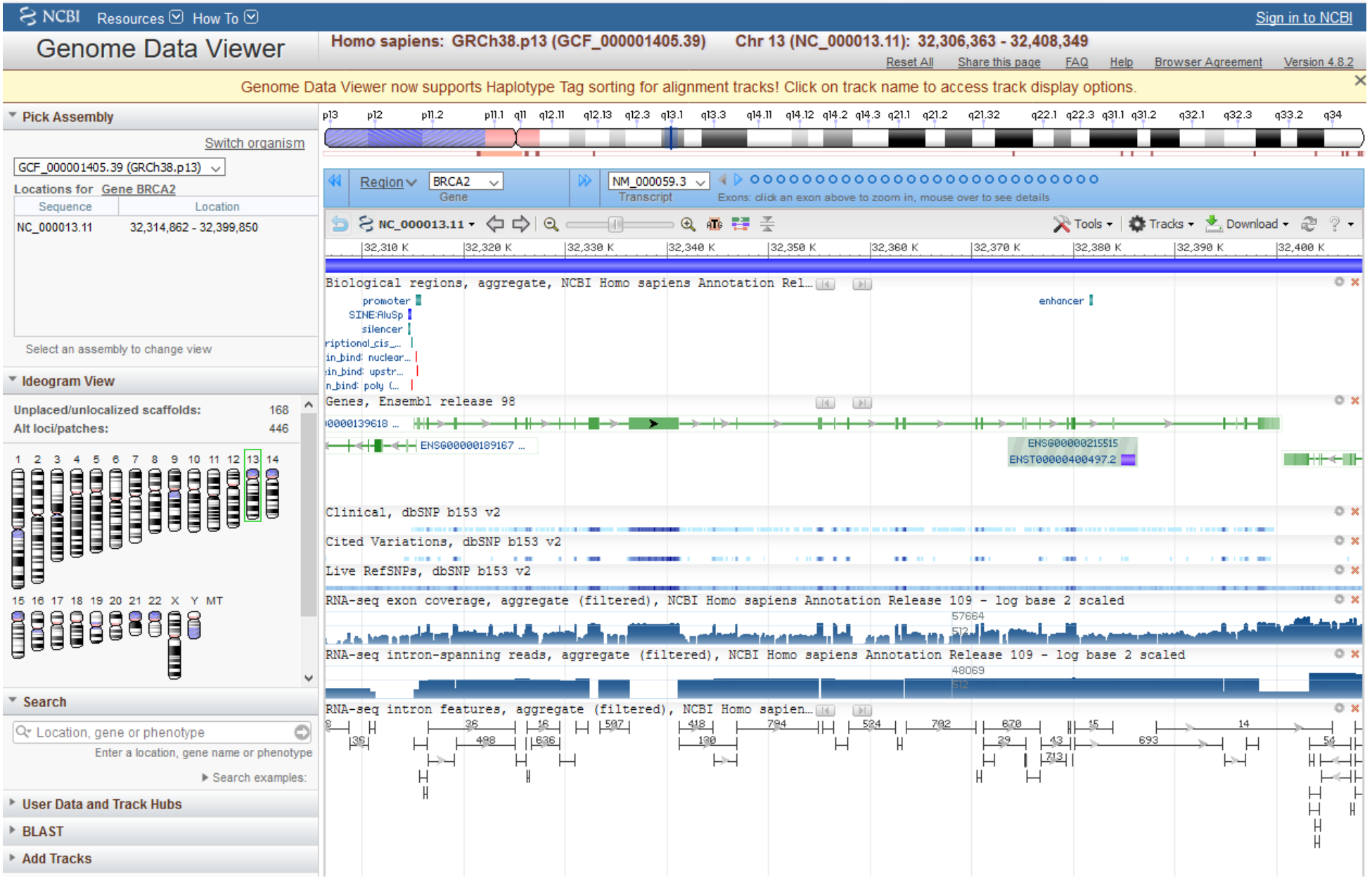
Recent activity

Turn Off Clear

BRCA2 HUMAN (2660) Nucleotide

BRCA2 (11289) Nucleotide

See more...


NCBI Resources How To
Sign in to NCBI
Genome Data Viewer
Homo sapiens: GRCh38.p13 (GCF_000001405.39) Chr 13 (NC_000013.11): 32,306,363 - 32,408,349
Reset All Share this page FAQ Help Browser Agreement Version 4.8.2
Genome Data Viewer now supports Haplotype Tag sorting for alignment tracks! Click on track name to access track display options.
Pick Assembly
Switch organism
GCF_000001405.39 (GRCh38.p13)
Locations for Gene BRCA2
Sequence Location
NC_000013.11 32,314,862 - 32,399,850
Select an assembly to change view
Ideogram View
Unplaced/unlocalized scaffolds: 168
Alt loci/patches: 446
Search
Location, gene or phenotype
Enter a location, gene name or phenotype
Search examples:
User Data and Track Hubs
BLAST
Add Tracks
Region BRCA2 Gene
NM_000059.3 Transcript
Exons: click an exon above to zoom in, mouse over to see details
NC_000013.11
Tools Tracks Download
Biological regions, aggregate, NCBI Homo sapiens Annotation Rel...
promoter
SINE:AluSp
silencer
enhancer
Genes, Ensembl release 98
ENSG00000189167 ...
ENSG00000215515
ENST00000400497.2
Clinical, dbSNP b153 v2
Cited Variations, dbSNP b153 v2
Live RefSNPs, dbSNP b153 v2
RNA-seq exon coverage, aggregate (filtered), NCBI Homo sapiens Annotation Release 109 - log base 2 scaled
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Homo sapiens Annotation Release 109 - log base 2 scaled
RNA-seq intron features, aggregate (filtered), NCBI Homo sapien...

NCBI Resources How To Sign in to NCBI

Nucleotide

Nucleotide BRCA2 HUMAN Search

Create alert Advanced Help

Species
Animals (1,776)
Plants (80)
Fungi (155)
Protists (65)
Bacteria (176)
Viruses (2)
Customize ...

Molecule types
genomic DNA/RNA (2,028)
mRNA (563)
Customize ...

Source databases
INSDC (GenBank) (1,994)
RefSeq (648)
Customize ...

Sequence Type
Nucleotide (2,624)
EST (35)
GSS (1)

Genetic compartments
Mitochondrion (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

GENE Was this helpful?

BRCA2 – BRCA2 DNA repair associated

Homo sapiens (human)

Also known as: BRCC2, BROVCA2, FACD, FAD, FAD1, FANCD, FANCD1, GLM3, PNCA2, XRCC11

GeneID: 675

RefSeq transcripts (1) RefSeq proteins (1) RefSeqGene (1) PubMed (1,766)

Orthologs Genome Browser BLAST Download

RefSeq Sequences +

Items: 1 to 20 of 2660

<< First < Prev Page 1 of 133 Next > Last >>

1. **Human** breast cancer susceptibility (**BRCA2**) mRNA, complete cds
10,987 bp linear mRNA
Accession: U43746.1 GI: 1161383
Protein PubMed Taxonomy
GenBank FASTA Graphics

2. H.sapiens **brca2** gene exon 2 (and joined coding region)
1,106 bp linear DNA
Accession: X95152.1 GI: 1171546
Protein PubMed Taxonomy
GenBank FASTA Graphics

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(1110)*
unidentified *(384)*
Escherichia coli *(163)*
Mus musculus *(83)*
Rattus norvegicus *(29)*
All other taxa *(891)*
More...

Find related data

Database: Select

Find items

Search details

```
BRCA2[All Fields] AND ("Homo
sapiens"[Organism] OR HUMAN[All
Fields])
```

Search See more...

Recent activity

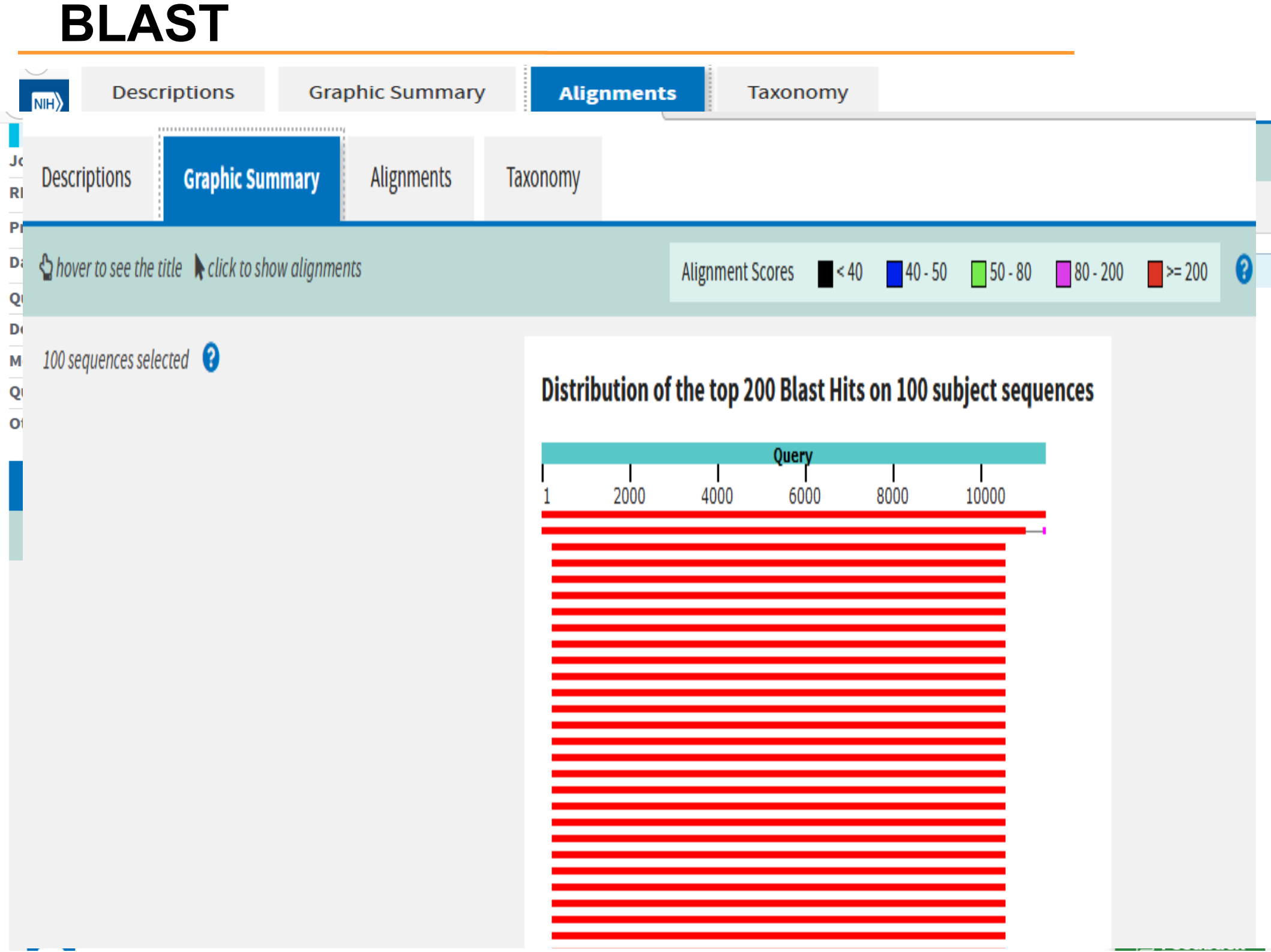
Turn Off Clear

BRCA2 HUMAN (2660) Nucleotide

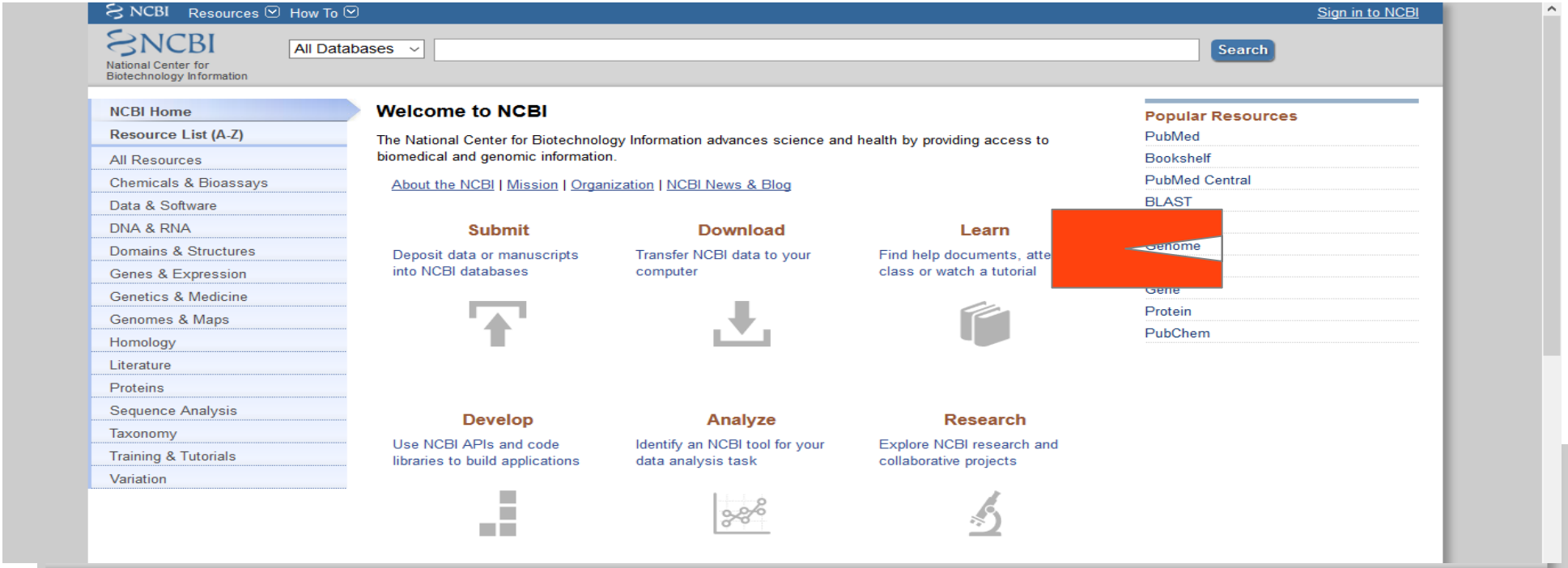
BRCA2 (11289) Nucleotide

See more...

BLAST


Descriptions
Graphic Summary
Alignments
Taxonomy
Descriptions
Graphic Summary
Alignments
Taxonomy
hover to see the title
click to show alignments
Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
100 sequences selected
Distribution of the top 200 Blast Hits on 100 subject sequences
Query
1
2000
4000
6000
8000
10000

NCBI


NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, atte class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build applications
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Genome
Gene
Protein
PubChem


ORGANELLE
Was this helpful?
Human mitochondrial reference genome
Homo sapiens
Included in the human reference assembly (GRCh38.p13)
RefSeq: NC_012920.1 Length: 16,569 bp circular
TRNF
RefSeq protein (13) PopSet (1,010) PubMed (2)
RNR1 RNR2 ND1 ND2 COX1 COX2 ATP6 COX3 ND4 ND5 ND6 CYTB
Genome Browser
Primer-BLAST
Download
See all archival Homo sapiens complete mitochondrial genomes (48,476)
Search details
("Homo sapiens"[Organism] OR Homo sapiens[All Fields]) AND mitochondrion[All Fields]
Search
See more...
Recent activity
Turn Off Clear
Homo sapiens mitochondrion (0)
Genome
Homo sapiens
Genome
homo sapiens[orgn] (1)
Genome

NCBI:Μεταλλάξεις/ Πολυμορφισμοί

Ορισμοί

SNPs: Πολυμορφισμοί του ενός νουκλεοτιδίου
c-SNPs: coding single nucleotide polymorphisms (Πολυμορφισμοί του ενός νουκλεοτιδίου σε cDNA ακολουθίες,δηλαδή σε αλληλουχίες που κωδικοποιούν αμινοξέα)
SAPs: Πολυμορφισμοί του ενός αμινοξέος

Μετάλλαξη με άλλο νόημα: -> SAP
Μετάλλαξη χωρίς νόημα: -> STOP
Προσθήκη / Έλλειψη νουκλεοτιδίων -> Μεταλλάξεις αλλαγής του αναγνωστικού πλαισίου (frameshift)…

NCBI: Επιπλέον...

Πρακτική Άσκηση

- Επισκεφθείτε την σελίδα του NCBI και με «query» την πρωτεϊνη «hsp1» από το φυτό Cistus creticus αναζητήστε με τον αλγόριθμο BLAST αλληλουχίες που μοιάζουν σε άνθρωπο (επιλογή-organism).

- Περιοριστείτε στα πρώτα 5 αποτελέσματα.

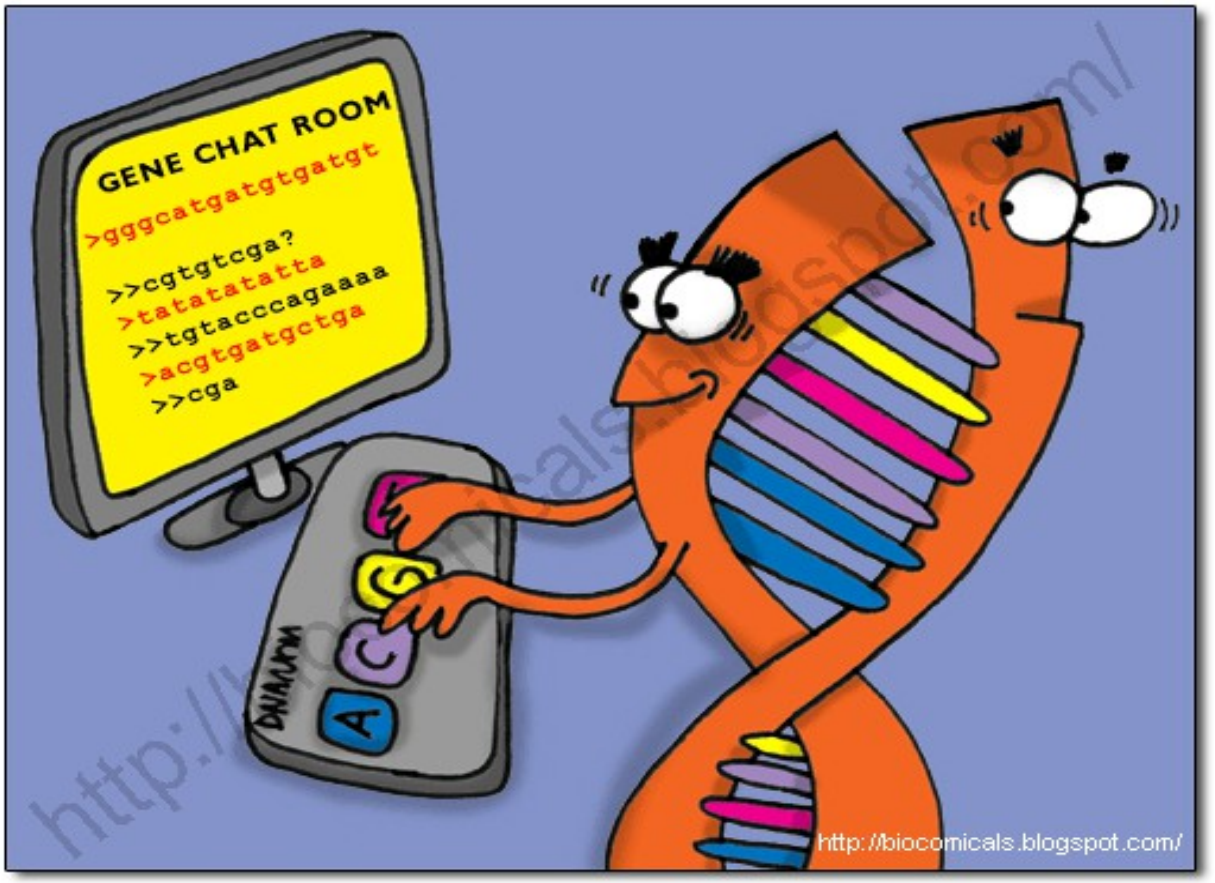
- Εξάγατε τις ακολουθίες των 5 πρωτεϊνών αυτών σε «fasta format» ή σημειώστε τα accession numbers τους.

- Με «query» αυτές τις αλληουχίες ή τα accession numbers τους επαναλλάβατε την αναζήτηση (μία μία εννοείται) ανάποδα δηλ. εναντίον των πρωτείνων των φυτών του γένους Cistus.
- Σε περίπτωση που δεν έχετε αποτελέσματα επεκτείνετε την αναζήτηση σε όλα τα θηλαστικά και σε όλα τα φυτά.

- **Σχολιάστε τα αποτελέσματά σας.**

- **Τι σχέση έχουν μεταξύ τους οι πρωτείνες που “χτυπούν¨ή μια πάνω στην άλλη και αντίστροφα με το BLAST?**

Ευχαριστώ



mkapasa@gmail.com