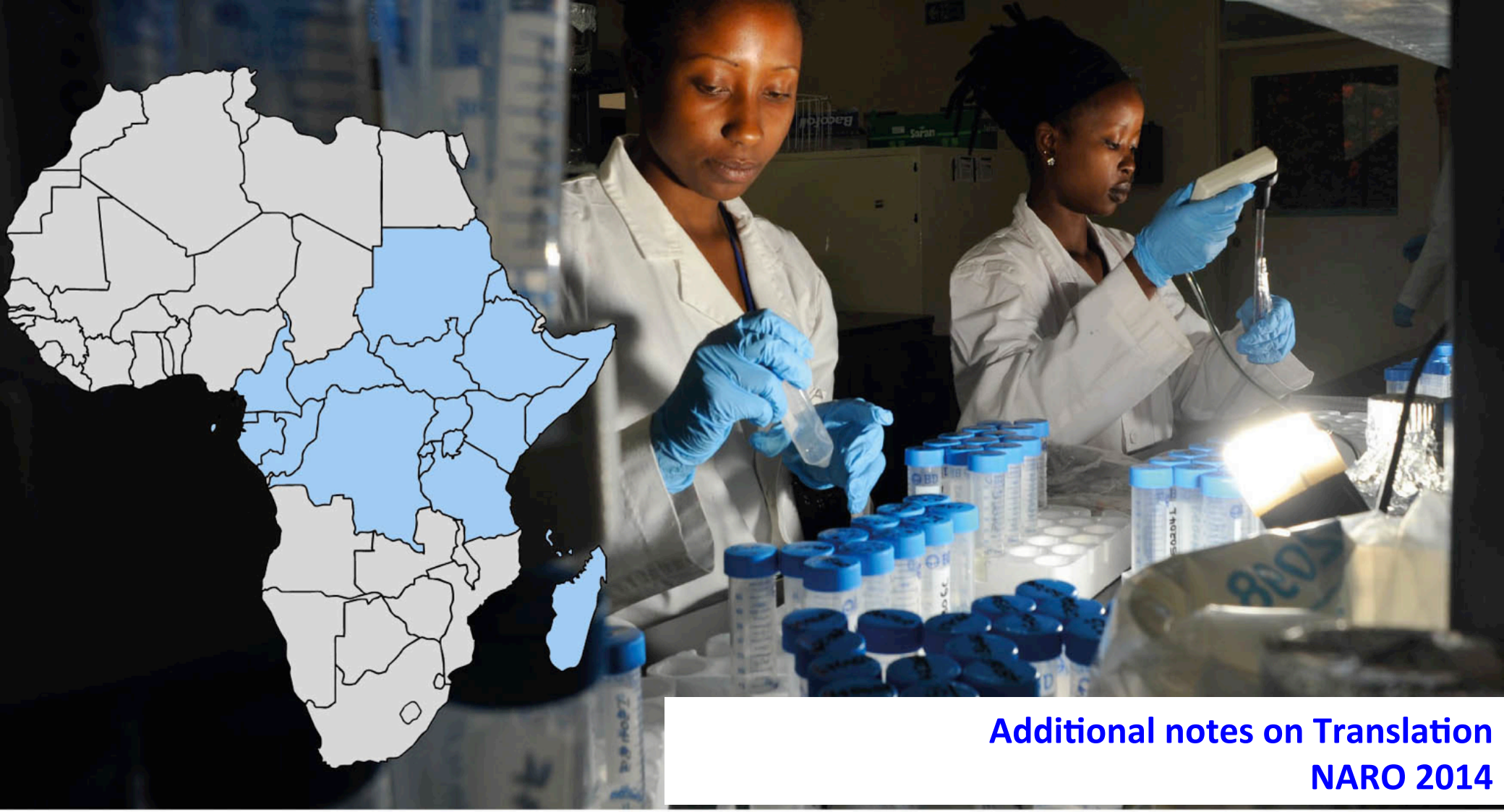


Additional notes on Translation
NARO 2014

Joyce Njoki Nzioki
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
http://www.Ilri.org/
j.n.njuguna@cgiar.org

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

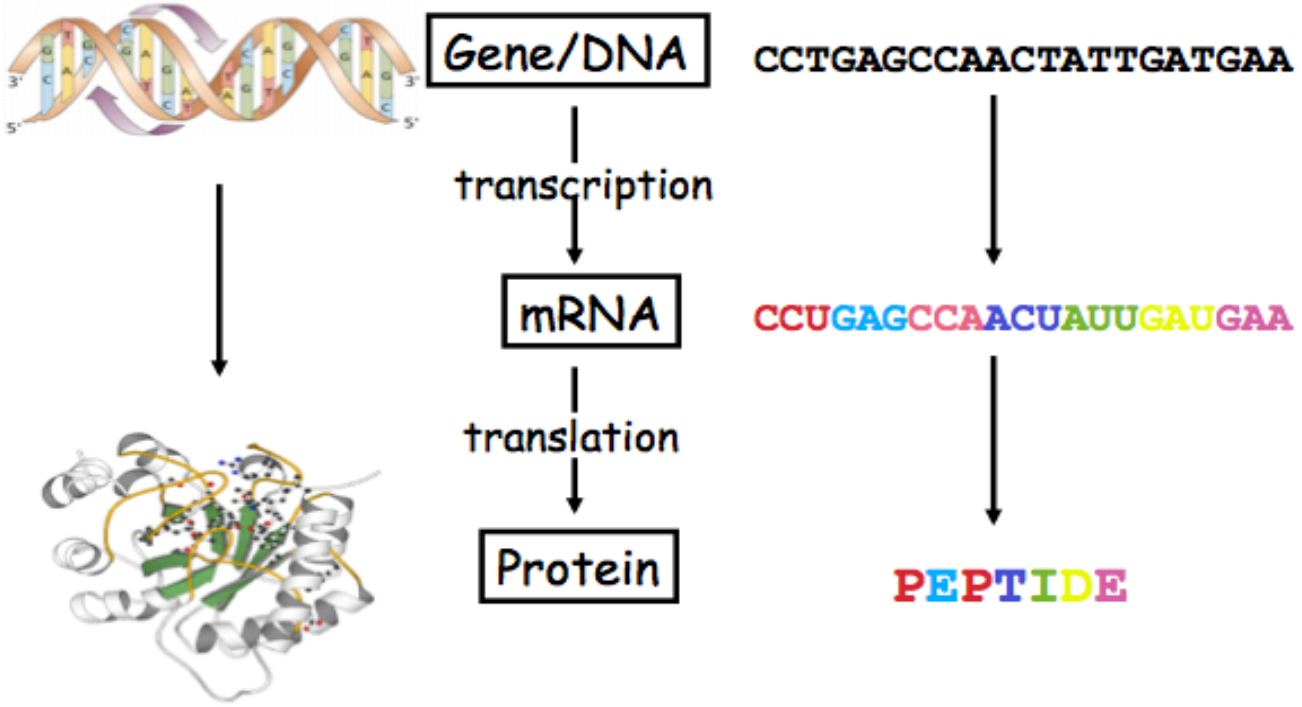
African Union

NEPAD
A PROGRAMME OF THE AFRICAN UNION

biosciences
eastern and central africa

Standard Genetic Code

- Genes are stretches of DNA that encode information for building proteins. “One gene encodes one protein”
- A triplet of nucleotides codes for an amino acid the building blocks of proteins



Standard Genetic Code

Glycine	(GLY)	GG*	
Alanine	(ALA)	GC*	
Valine	(VAL)	GT*	
Leucine	(LEU)	CT*	
Isoleucine	(ILE)	AT(*-G)	
Serine	(SER)	AGT	AGC
Threonine	(THR)	AC*	
Aspartic acid	(ASP)	GAT	GAC
Glutamic acid	(GLU)	GAA	GAG
Lysine	(LYS)	AAA	AAG
Start	ATG	CTG	GTG

Arginine	(ARG)	CG*	
Asparagine	(ASN)	AAT	AAC
Glutamine	(GLN)	CAA	CAG
Cysteine	(CYS)	TGT	TGC
Methionine	(MET)	ATG	
Phenylalanine	(PHE)	TTT	TTC
Tyrosine	(TRY)	TAT	TAC
Tryptophan	(TRP)	TGG	
Histidine	(HIS)	CAT	CAC
Proline	(PRO)	CC*	
Stop	TGA	TAA	TAG

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

NEPAD
A PROGRAMME OF THE AFRICAN UNION

biosciences
eastern and central africa

Translating nucleotide sequences

- DNA codes for amino acids a three letter genetic code.
- Translation of DNA to Amino acids can be done in 6 different reading frames.

GATTCGTACG
CTAAGCATGC

GATTCGTACG

1. GAT (Asp) TCG (Ser) TAC (Thr)
2. G ATT (Ile) CGT (Arg) ACG (Thr)
3. GA TTC (Phe) GTC (Val) CG

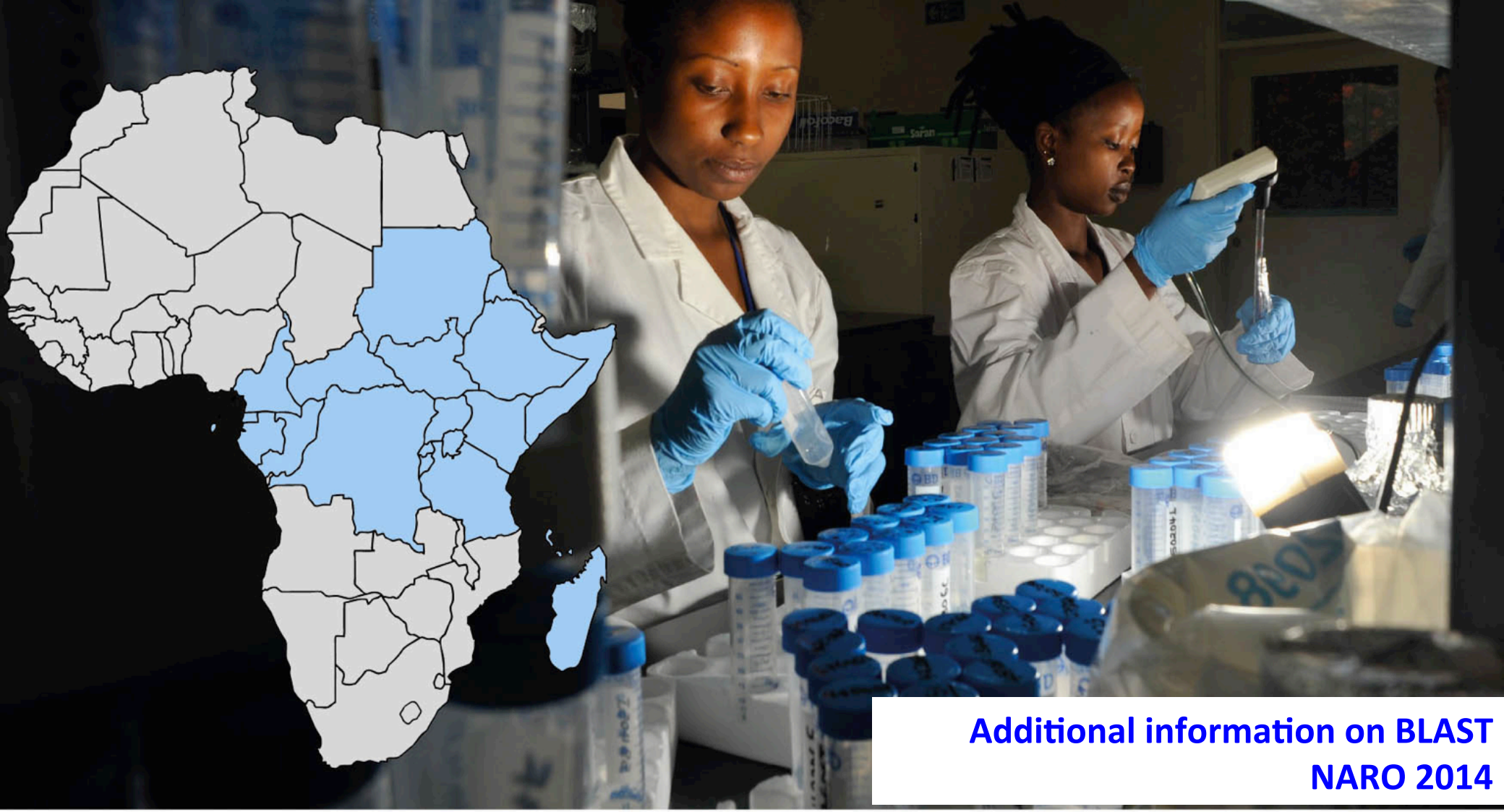
CTAAGCATGC

4. CTA (Leu) AGC (Ser) ATG (Met) C
5. C TAA (#) GCA (Ala) TGC (Arg)
6. CT AAG (Lys) CAT (His) GC

Translation tools

- There various computational tools available for sequence translation:

1. CLC main workbench: (Practicals in the next session).
2. Expasy translation tool: http://www.expasy.ch/tools/dna.html
3. EMBOSS: transeq(translate nucleic acids)
4. SeWeR analysis: (Sequence analysis using web resources) http://www.bioinformatics.org/SeWeR/



Additional information on BLAST
NARO 2014

Joyce Njoki Nzioki
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
http://www.Ilri.org/
j.n.njuguna@cgiar.org

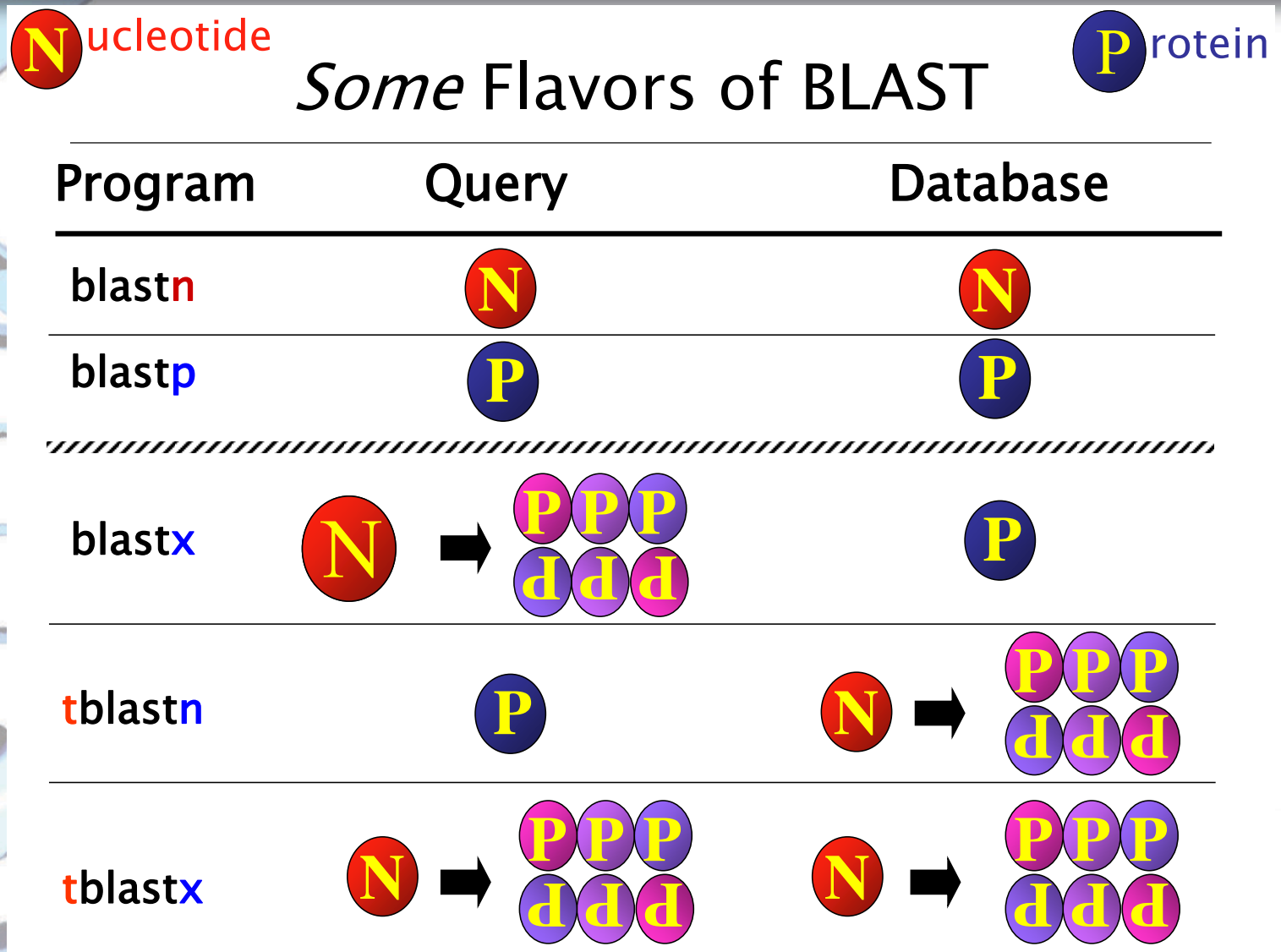
ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

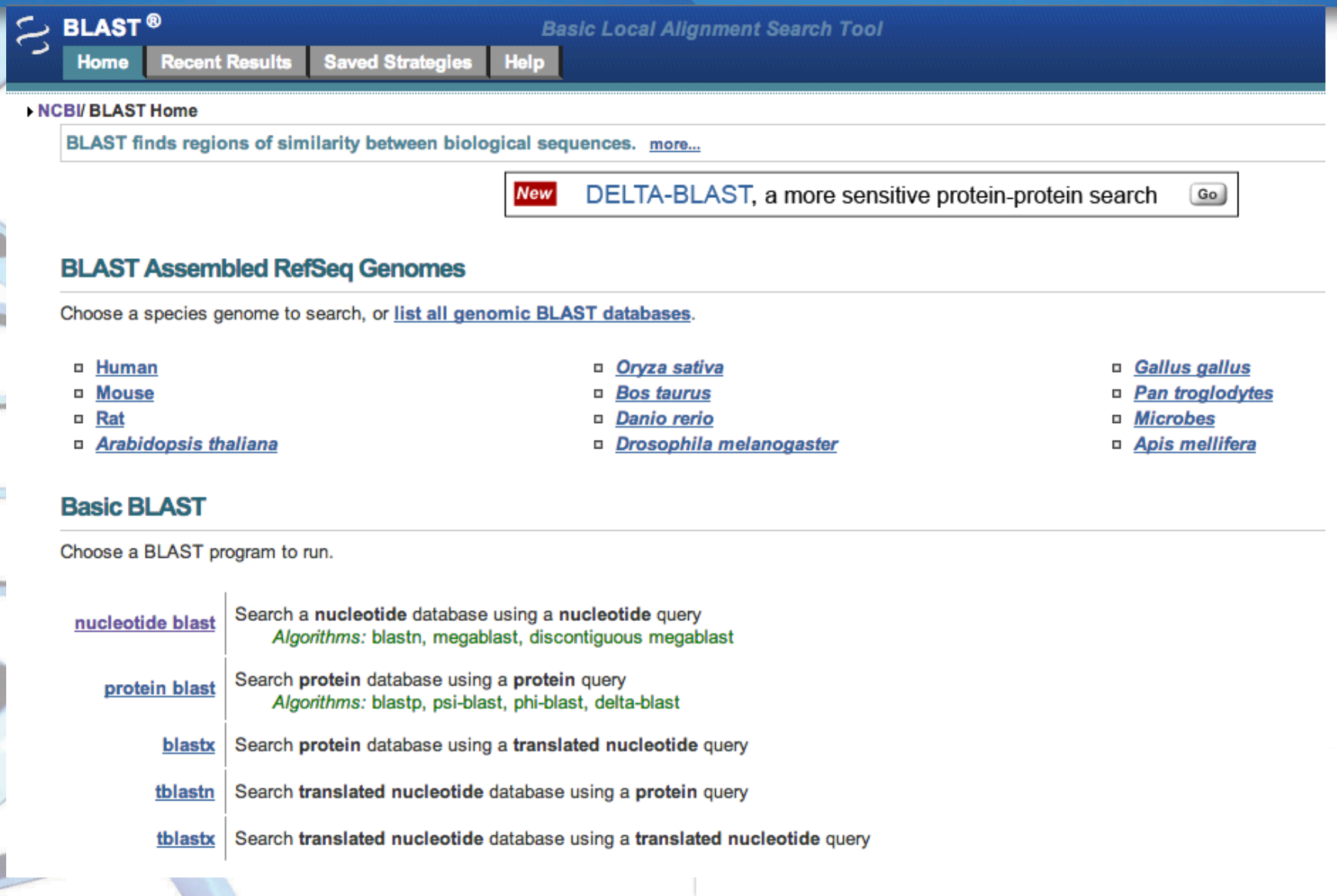
NEPAD
A PROGRAMME OF THE AFRICAN UNION

biosciences
eastern and central africa

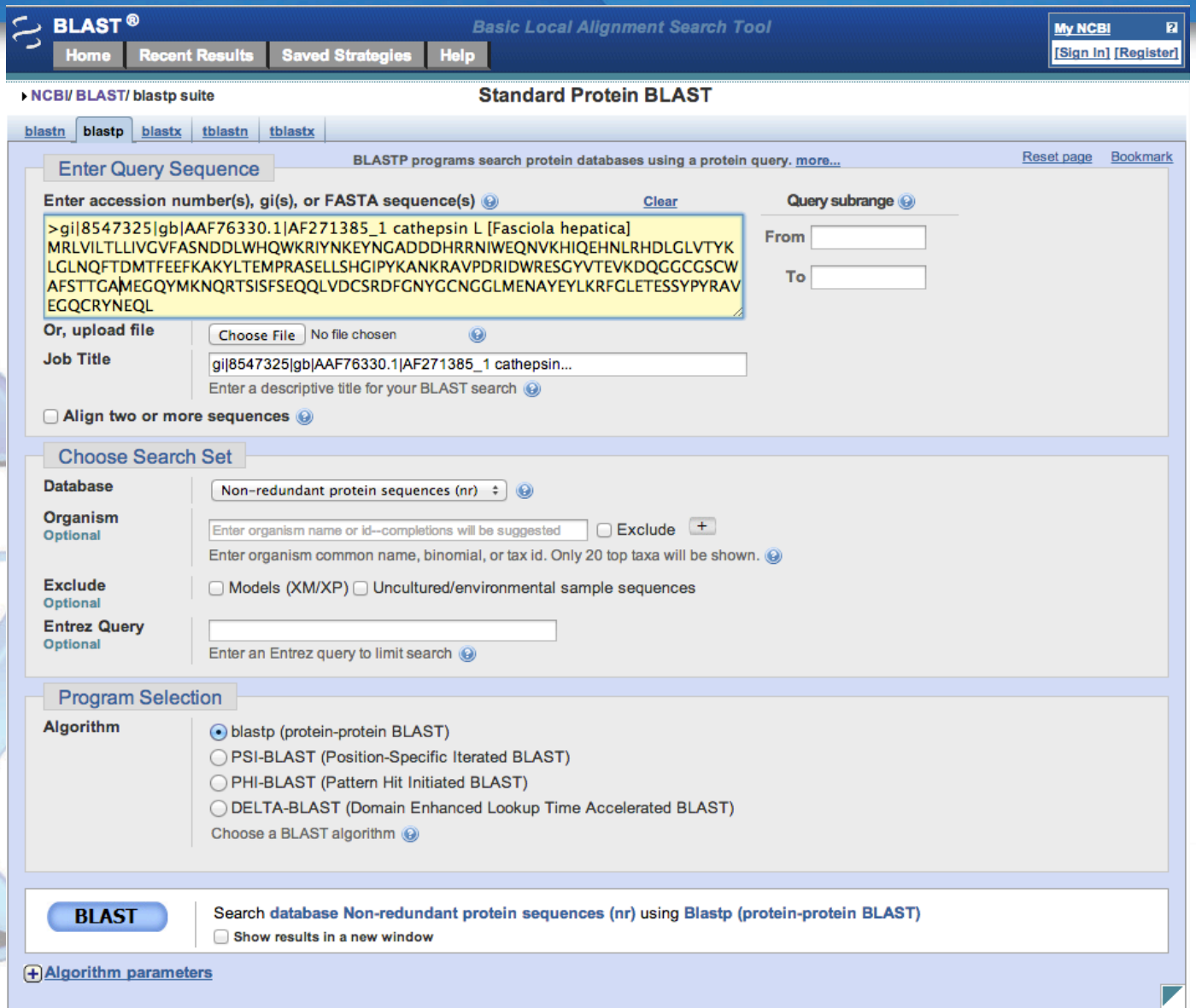
BLAST Flavors



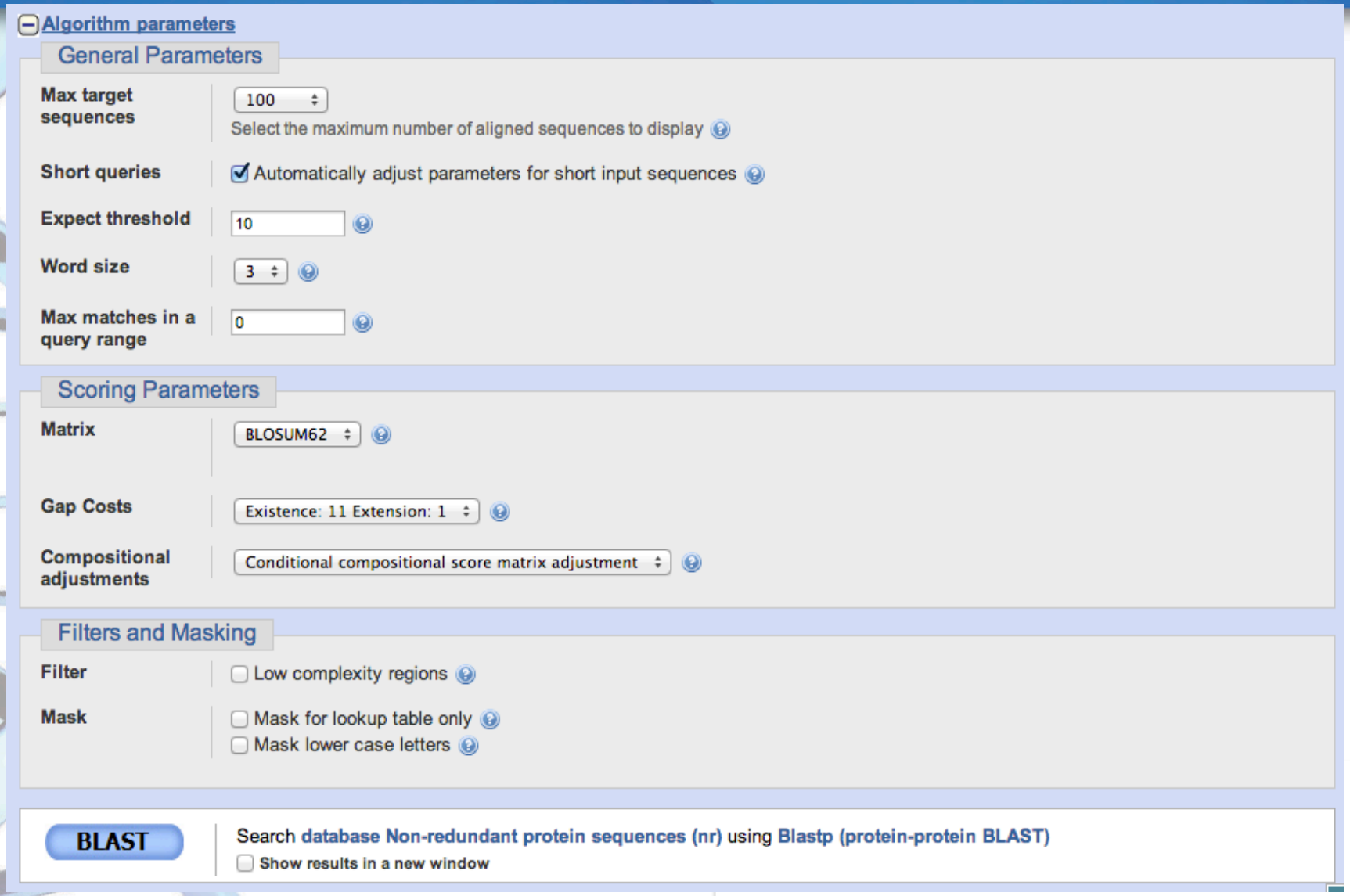
NCBI BLAST


BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New
DELTA-BLAST, a more sensitive protein-protein search
Go
BLAST Assembled RefSeq Genomes
Choose a species genome to search, or list all genomic BLAST databases.
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query

BLASTp


BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>gi|8547325|gb|AAF76330.1|AF271385_1 cathepsin L [Fasciola hepatica]
MRLVILTLLIVGVFASNDDLWHQWKRIYNKEYNGADDDHRRNIWEQNVKHIQEHNLRHDLGLVTYK
LGLNQFTDMTFEEFKAKYLTEMPRASELLSHGIPYKANKRAVPDRIDWRESGYVTEVKDQGGCGSCW
AFSTTGAMEGQYMKNQRTSISFSEQQLVDCSRDFGNYGCNGGLMENAYEYLKRFGLETESSYPYRAV
EGQCRYNEQL
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
gi|8547325|gb|AAF76330.1|AF271385_1 cathepsin...
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters

Advanced BLAST Parameters



BLAST Results

gi|8547325|gb|AAF76330.1|AF271385_1 cathepsin...

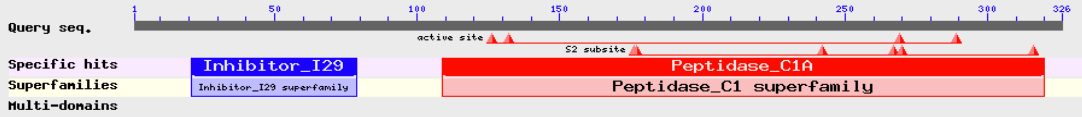
Query ID	lcl|9878	**Database Name**	nr
Description	gi|8547325|gb|AAF76330.1|AF271385_1 cathepsin L [Fasciola hepatica]	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	326	**Program**	BLASTP 2.2.27+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

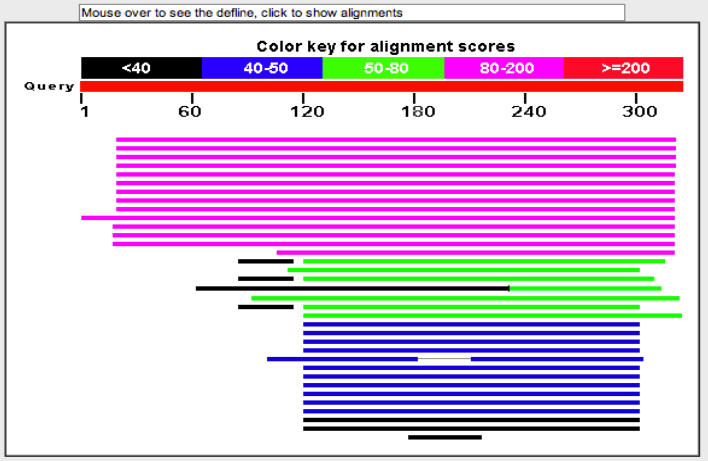
Graphic Summary

Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 40 Blast Hits on the Query Sequence



Descriptions

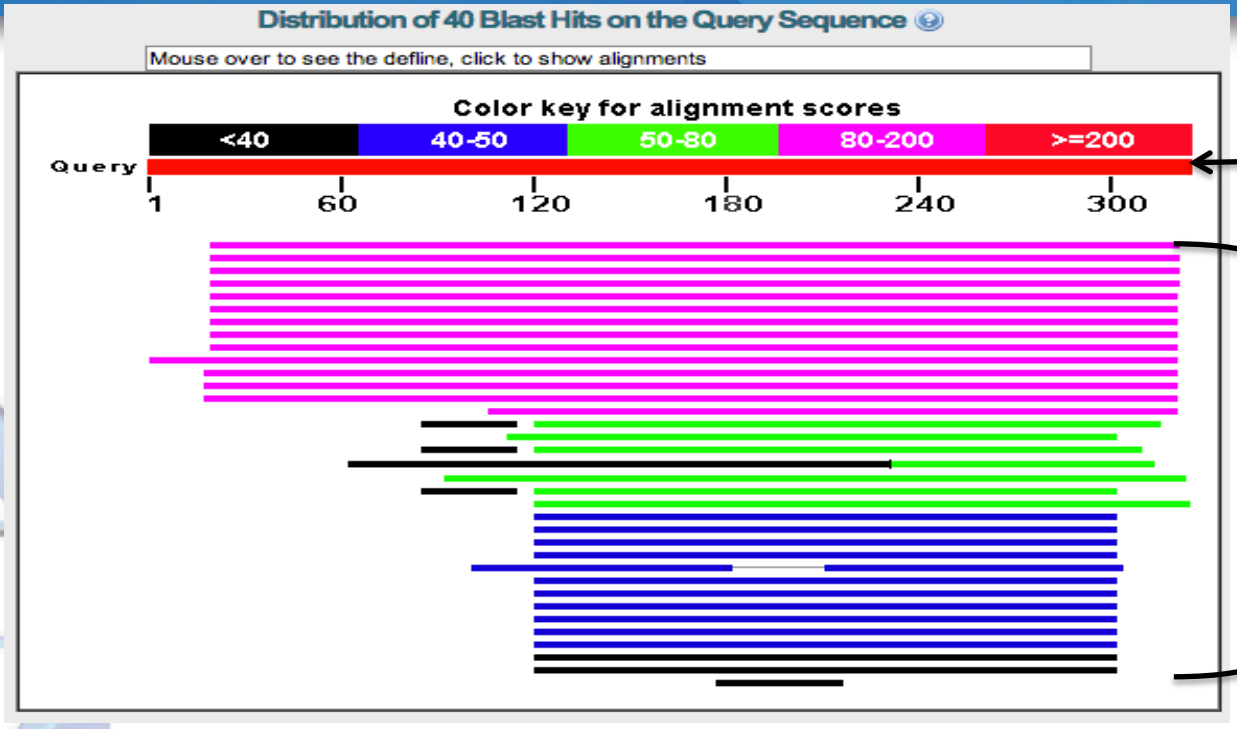
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Max ident	Accession
vivapain-3 [Plasmodium vivax] >gb|AAT36276.1| vivapain-3 [Plasmodium vivax] >gb|AAT36277.1| vivapain-3	171	171	92%	1e-48	35%	AAT36272.1
vivapain-3 [Plasmodium vivax] >gb|AAT36266.1| vivapain-3 [Plasmodium vivax] >gb|AAT36267.1| vivapain-3	171	171	92%	1e-48	35%	AAT36265.1

Graphical BLAST Results



Query Sequence

Blast Hits,
A mouse over gives you the details.
Click to view alignment

- This is a graphical view of the distribution of BLAST hits on the query sequence.
- The length of the hits shows the query coverage and region of similarity.
- The colors represent similarity scores with red been the highest down to black

Hit List BLAST results

- *This gives the names of sequences similar to your query sequence ranked by similarity*

Bit score values < 50 unreliable

% Query coverage

E-value

Description	Max score	Total score	Query cover	E value	Max ident	Accession
vivapain-1 [Plasmodium vivax Sal-1] >gi|148801225|gb|EDL42630.1| vivapain-1 [Plasmodium vivax] >gb|EDL42630.1|	1306	1306	100%	0.0		XP_001612423.1
cysteine proteinase precursor [Plasmodium vivax Sal-1] >gb|AAT36221.1| vivapain-1 [Plasmodium vivax] >gb|AAT36	1203	1203	92%	0.0	100%	XP_001615807.1
RecName: Full=Cysteine proteinase; Flags: Precursor >gb|AAA60368.1| cysteine proteinase [Plasmodium vivax]	1202	1202	92%	0.0	99%	P42666.1
cysteine proteinase precursor [Plasmodium cynomolgi strain B] >dbj|GAB68005.1| cysteine proteinase precursor [Plas	963	963	92%	0.0	82%	XP_004223952.1
trophozoite cysteine proteinase precursor [Plasmodium knowlesi strain H] >emb|CAQ41558.1| trophozoite cysteine pr	897	897	92%	0.0	77%	XP_002260291.1
vivapain-1 [Plasmodium vivax Sal-1] >gb|EDL42565.1| vivapain-1 [Plasmodium vivax]	662	662	49%	0.0	100%	XP_001612358.1
cysteine proteinase, partial [Plasmodium cynomolgi]	632	632	52%	0.0	91%	AAC47033.1
vivapain-1 [Plasmodium vivax Sal-1] >gb|EDL42547.1| vivapain-1 [Plasmodium vivax]	580	580	45%	0.0	100%	XP_001612340.1
cysteine proteinase, partial [Plasmodium fragile]	582	582	52%	0.0	83%	AAC47034.1
RecName: Full=Trophozoite cysteine proteinase; Short=TCP; Flags: Precursor >gb|AAA29578.1| cysteine proteinase	468	468	56%	3e-154	63%	P25805.1
cysteine proteinase falcipain-1 [Plasmodium falciparum 3D7] >gb|AAN37166.1| cysteine proteinase falcipain-1 [Plasm	467	467	56%	3e-154	63%	XP_001348727.1
trophozoite cysteine proteinase precursor [Plasmodium berghei strain ANKA] >emb|CAH94555.1| trophozoite cysteine	432	432	92%	2e-141	42%	XP_677643.1

Sort by E value

Sequence Definition, click to view the pairwise alignment

Accession number, Link to the record in NCBI Entrez

Pairwise alignment results

Score	Expect	Method	Identities	Positives	Gaps
167 bits(424)	2e-47	Compositional matrix adjust.	110/330(33%)	168/330(50%)	38/330(11%)

```
Query  21   WHQWKRIYNKEYNGADDDHRRNI-WEQNVKHIQEHNLRHDLGLVTYKLGLNQFTDMTFEE  79
            ++ + + Y ++Y   ++  +R + + +N++ I+ HN R +   V Y+ G+NQF D++F E
Sbjct  167  FYLFVKEYGRKYKTEEEMQQRYLAFVENLEKIKAHNSREN---VLYRKGMNQFGDLSFGE  223

Query  80   FKAKYLTEMPRASELLSHGIPYKANKRAVPDRI----------DWRESGYVTEVKDQG  127
            FK KYLT      +     +    N   V D+              DWR    VT VKDQ
Sbjct  224  FKKKYLTLKSFDFKTFGGKLKRITNYEDVIDKYKPKDATFDHASYDWRLHKGVTPVKDQA  283

Query  128  GCGSCWAFSTTGAMEGQYMKNQRTSISFSEQQLVDCSRDFGNYGCNGGLMENAYE-YLKR  186
             CGSCWAFST G +E QY   +   +S SEQQ+VDCS    N GC GG +  A+E  ++
Sbjct  284  NCGSCWAFSTVGVVESQYAIRKNQLVSISEQQMVDCSTQ--NTGCYGGFIPLAFEDMIEM  341

Query  187  FGLETESSYPYRA-VEGQCRYNEQLGVAKVTGYYTVHSGDEVELQNLVGAEGPAAVALDV  245
             GL +   YPY A +   C+++      K+  +  +    E + +  +   GP +V++ V
Sbjct  342  GGLCSSEDYPYVADIPEMCKFDICEQKYKINNFLEI---PEDKFKEAIRFLGPLSVSIAV  398

Query  246  ESDFMMYRSGIYQSQTCSPDRLNHGVLAVGYGIQDGTD----------YWIVKNSWGTWW  295
              DF  YR GI+  + C  +  NH V+ VG+G +D  D           Y+IVKNSWG  W
Sbjct  399  SDDFAFYRGGIFDGE-CG-EAPNHAVILVGFGAEDAYDFDTKTMKKRYYYIVKNSWGVSW  456

Query  296  GEDGYIRM---VRKRGNMCGIASLASVPMV  322
            GE G+IR+   +      C + + A V +V
Sbjct  457  GEKGFIRLETDINGYRKPCSLGTEALVALV  486
```

Sbjct: segment from the hit sequence

Query: segment from the query sequence

The middle line is the consensus sequence.

BLAST Result Interpretation

- How do you make your conclusion on homology:
- E-value = Expected value. (this indicates the probability that the blast hit may have occurred by random chance).
- The lower the E-value (or the closer it is to 0) the more significant the hit. To be certain of homology your E-value must be below 10^{-4} or 0.001.
- % identity the higher the identity the increasing likelihood of homology.
- Query coverage – if a hit has high query coverage and similarity in increases the chances of homology.

Summary - for nucleotide sequences

Length	Database	Purpose	BLAST Program
20 bp or longer	Nucleotide	Identify the query sequence	blastn megablast
		Find similar nucleotide sequence.	blastn
		Find similar proteins to translated query in a translated nucleotide database	tblastx
	Protein	Find proteins coded in my query DNA sequence	blastx

Summary - for protein sequences

Length	Database	Purpose	BLAST Program
15 residues or Longer	Protein	Identify your query sequence or find protein sequences similar to it	blastp
		Find members of a protein family or build a custom position specific scoring matrix(PSSMs)	PSI-blast
		Find proteins similar to the query around a given pattern	PHI-blast
	Conserved domains	Find conserved domains in your query and identify other proteins with similar domains	CD-search
	Nucleic	Find similar sequences in a translated nucleotide sequence database.	tblastn