Mon, Oct 5

Lecture 4A:

Molecular sequences

Asimina triloba
© Matilda Adams/
Missouri Botanical Garden

Practical Bioinformatics (Biol 4220)
Instructor: Michael Landis
Email: michael.landis@wustl.edu

Lecture 4A outline

1. Molecular sequence features
2. Molecular sequence download
3. Some commands
4. Lab 4A overview

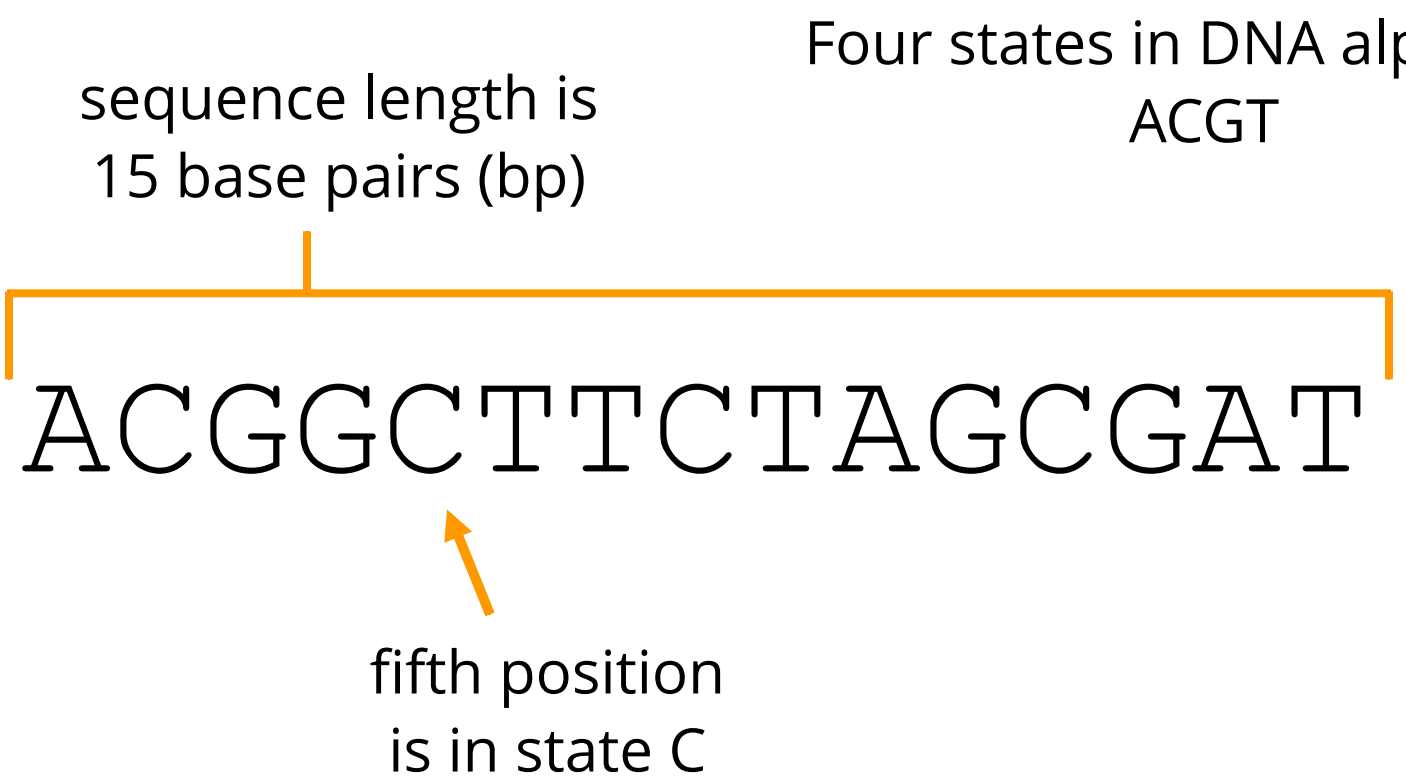
A ***molecular sequence*** is a linear sequence from a molecular alphabet

DNA sequences are most widely used molecular sequences; *RNA sequences* and *amino acid (AA) sequences* are not uncommon

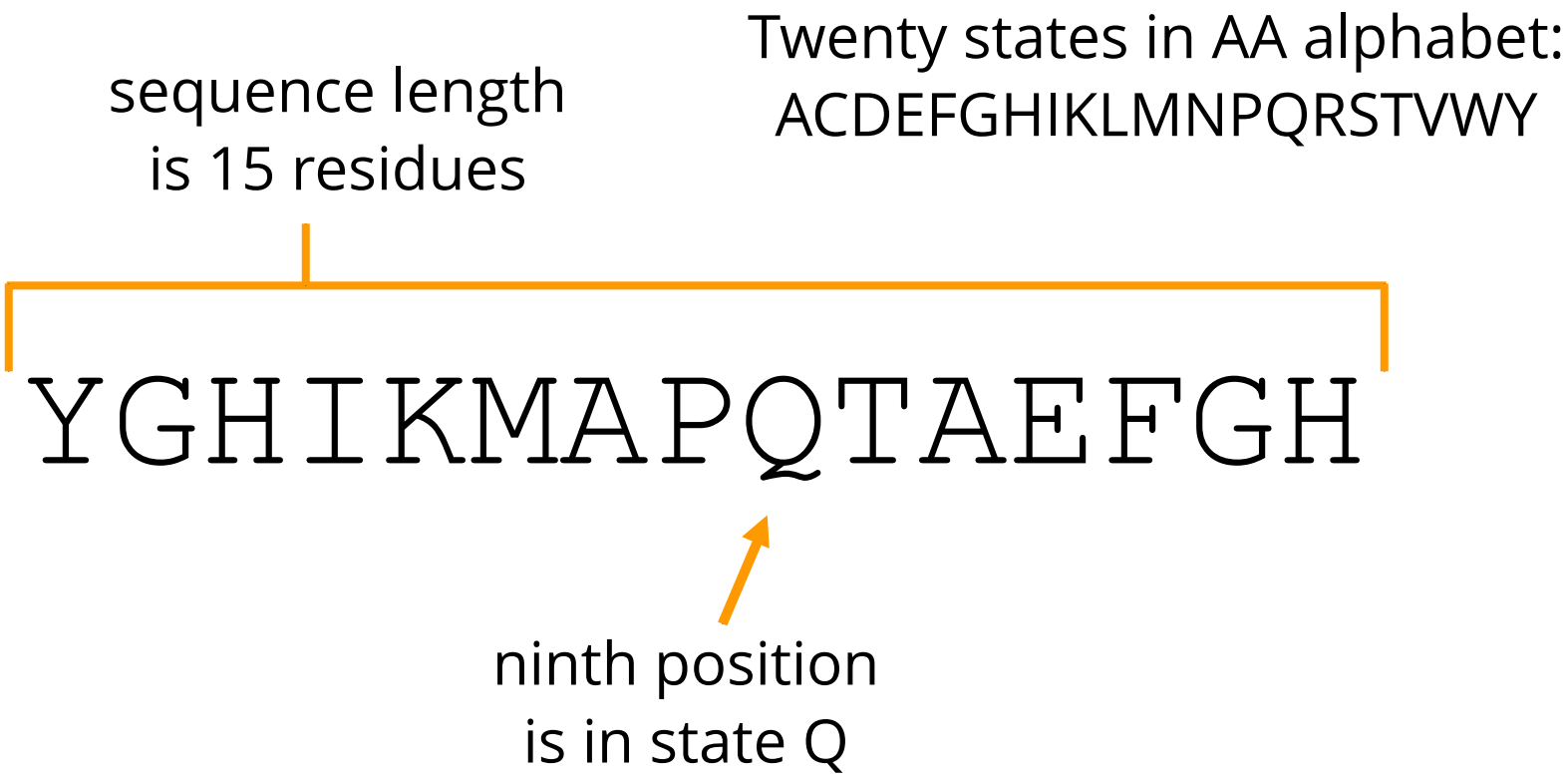
Sequence data informs all aspects of biology, *e.g.*

- mechanism for disease disease
- gene expression and development
- heredity and ancestry
- protein and cell function
- biodiversity surveys

Example DNA sequence



Example amino acid (AA) sequence



Genome

A genome is the complete set of sequences for a single sample (species, individual, tissue)

Genomic data has a *hierarchical structure*

- ***samples*** from a population have different genomes
- each genome is broken into multiple molecules, called ***chromosomes***
- each chromosome may contain some number of ***genes***
- each gene contains ***molecular characters***, and vary in total length and content

Genes

Genes are discrete heritable subunits of the genome

Some parts of the genome form natural subunits; others are named by researchers

Common genes (or loci)

- protein-coding sequences (exons, introns)
- transcription factor binding sites
- repeat regions (microsatellites)
- transposons, retrotransposons
- many loci have no known function

DNA to mRNA to AA

...ATGCGACGATGGATACCATAG...

DNA to mRNA to AA

...ATGCGACGATGGATACCATAG...

1. DNA *transcribed* into mRNA

AUGCGACGAUGGAUACCAUAG

DNA to mRNA to AA

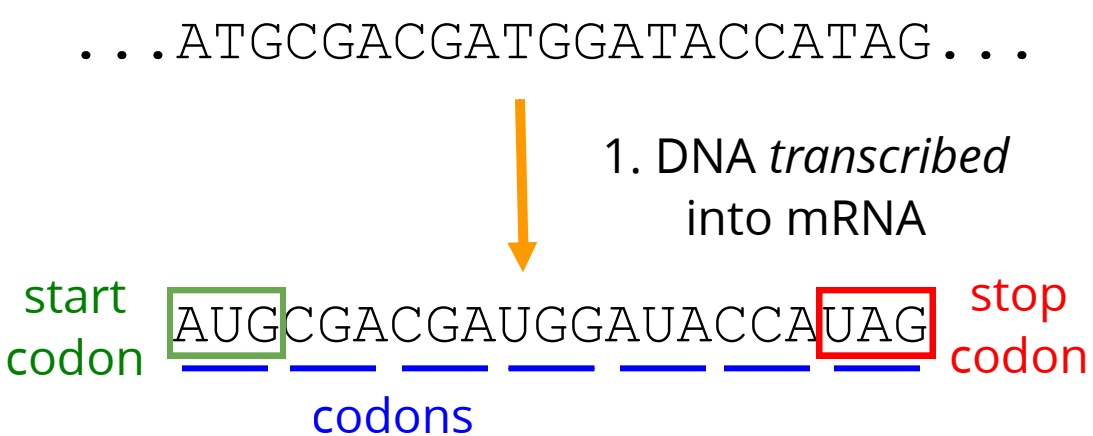
...ATGCGACGATGGATACCATAG...

1. DNA *transcribed* into mRNA

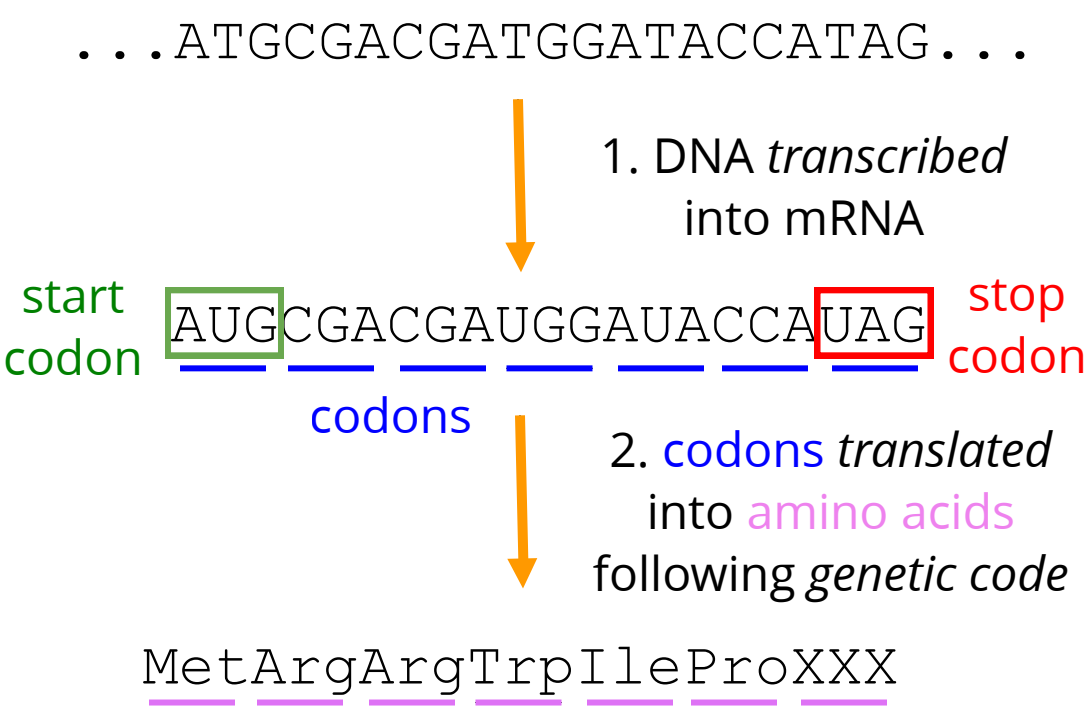
AUGCGACGAUGGAUACCAUAG

codons

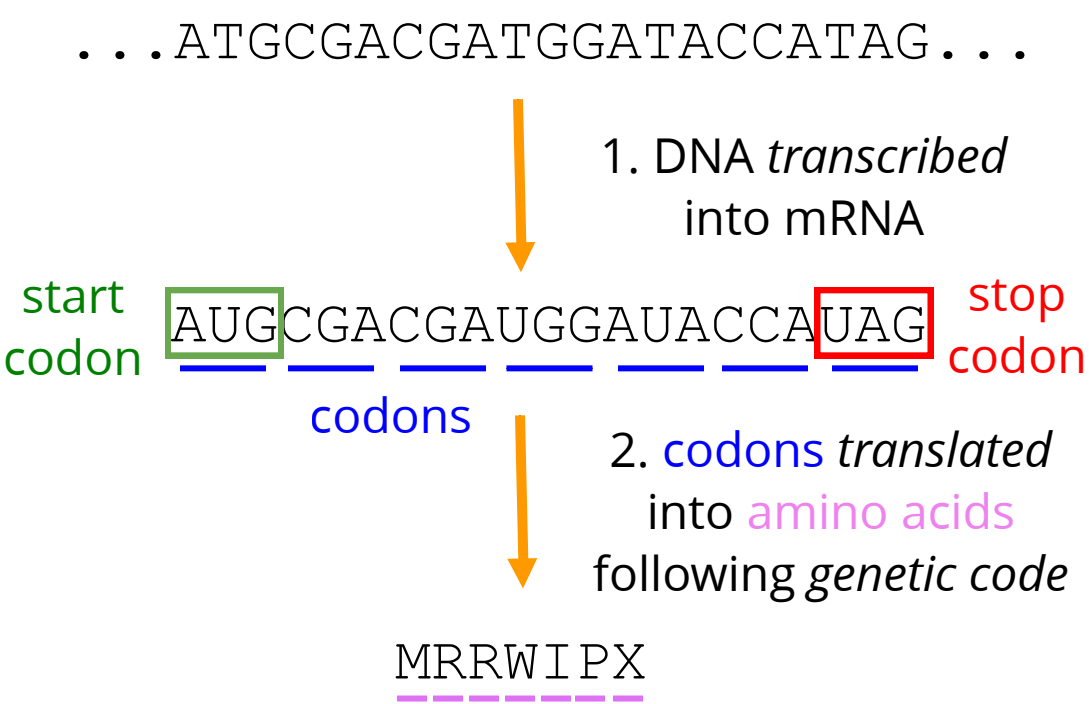
DNA to mRNA to AA



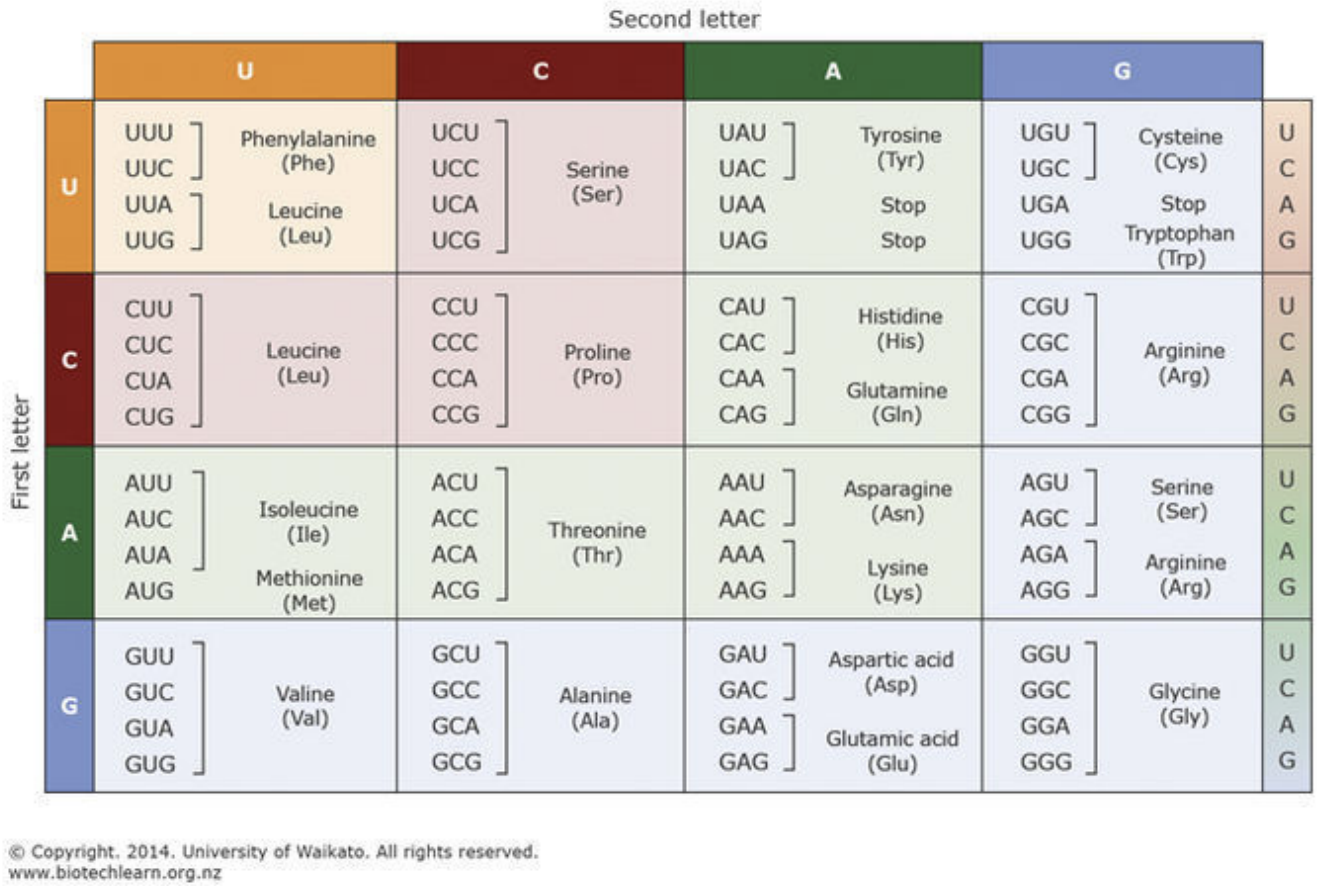
DNA to mRNA to AA



DNA to mRNA to AA



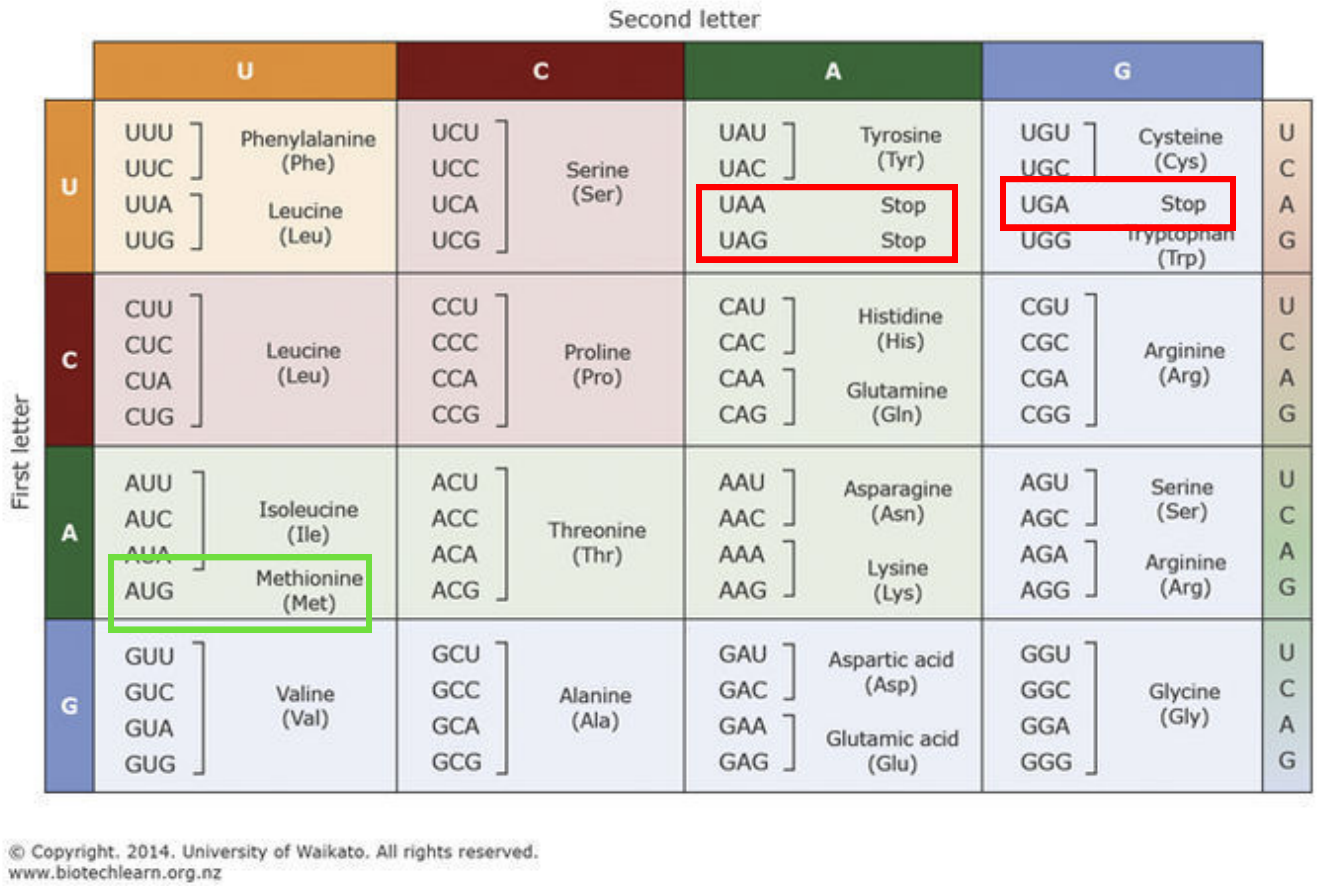
Genetic code



© Copyright. 2014. University of Waikato. All rights reserved.
www.biotechlearn.org.nz

Codons determine AA translation

Genetic code

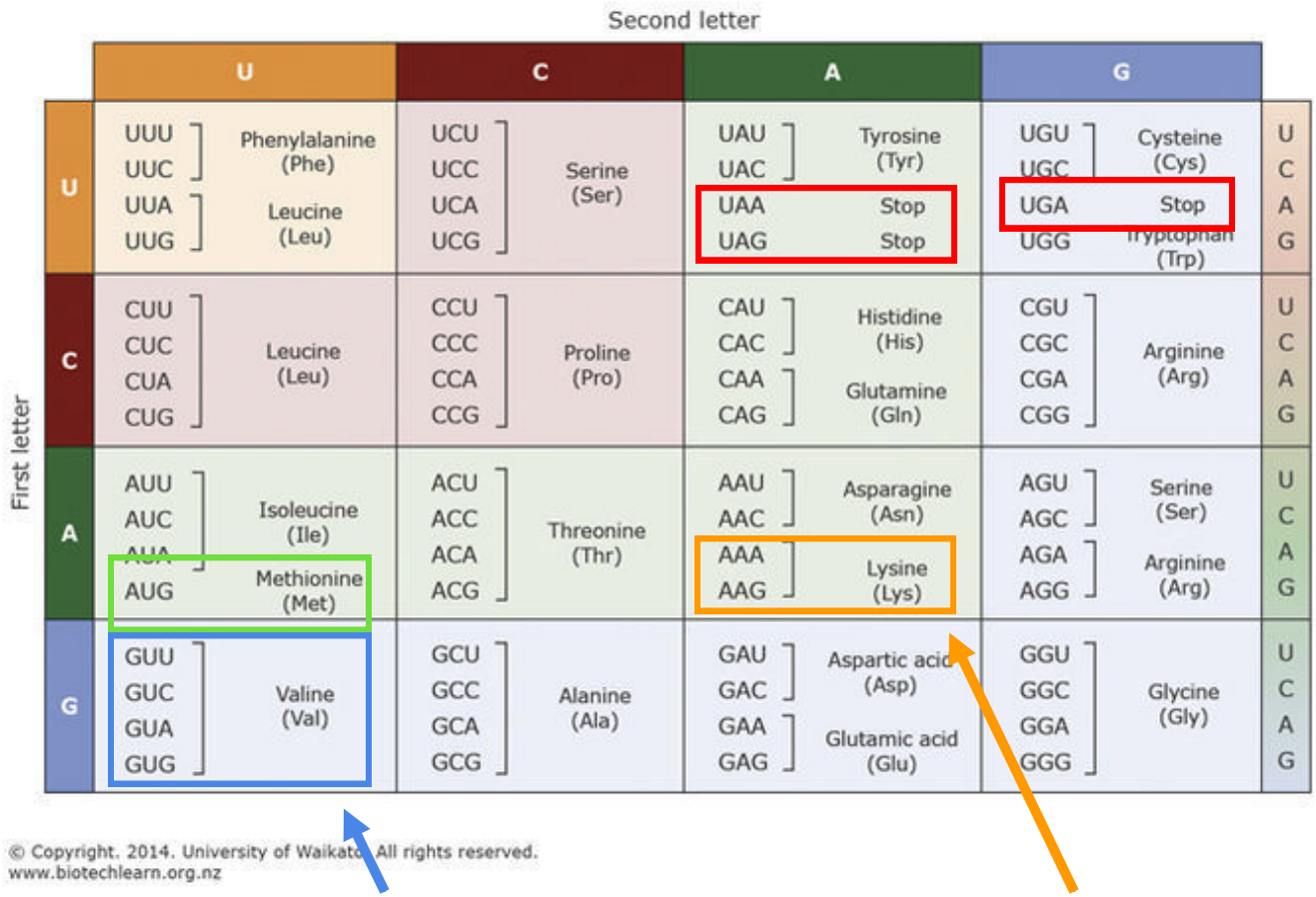


© Copyright. 2014. University of Waikato. All rights reserved.
www.biotechlearn.org.nz

start codon

stop codons

Genetic code



stop codons

start codon

GUU, GUC, GUA, and GUG all encode Valine

AAA and AAG encode Lysine

Protein structure

Amino acid sequences

Primary Structure = sequence of amino acids

3-letter code

Lys-Thr-Tyr-Phe-Pro-His-Phe-Asp-Leu-Ser-His-**Gly** ...

1-letter code

KTYFPHFDLSH**G**

Protein structure

Amino acid sequences

Primary Structure = sequence of amino acids

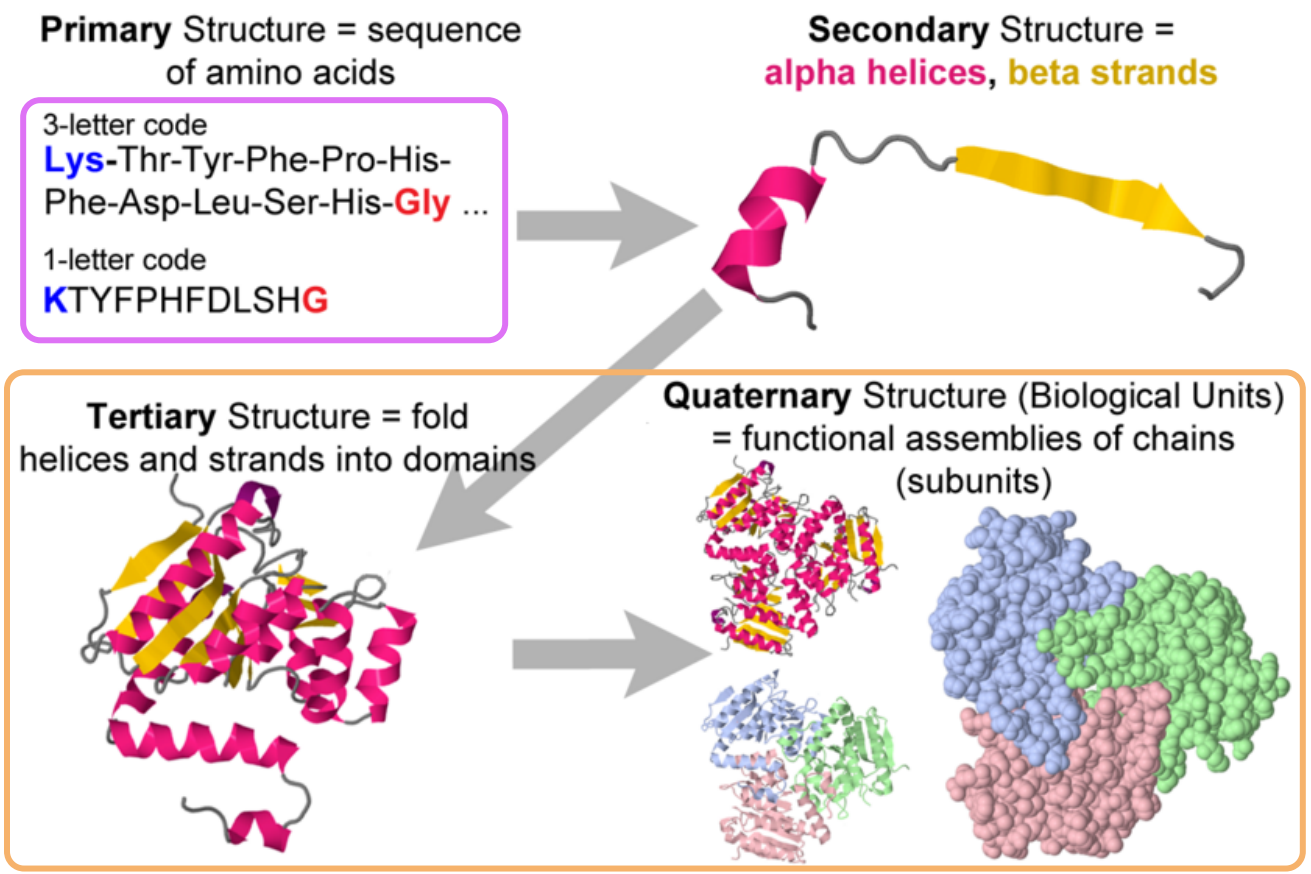
3-letter code
Lys-Thr-Tyr-Phe-Pro-His-Phe-Asp-Leu-Ser-His-**Gly** ...

1-letter code
KTYFPHFDLSH**G**

Secondary Structure = **alpha helices**, **beta strands**

Protein structure

Amino acid sequences

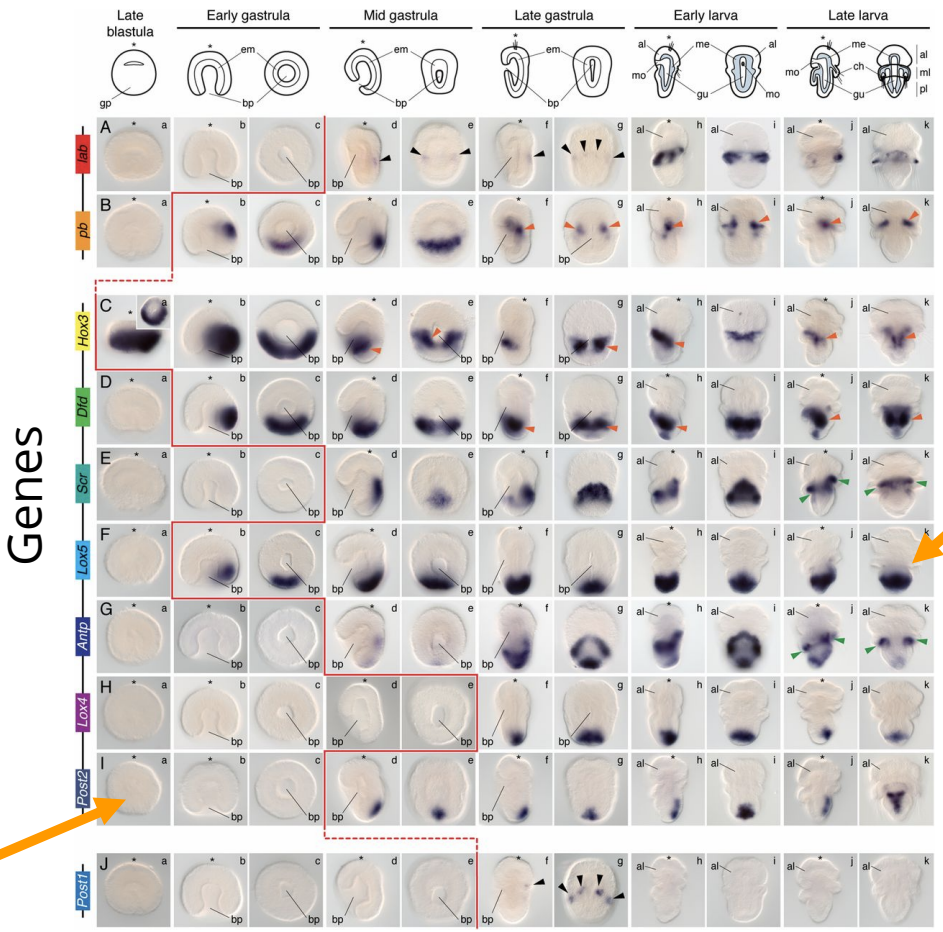


higher order protein structure
increasingly determines
protein function

proteopedia.org/wiki/index.php/Protein_primary%2C_secondary%2C_tertiary_and_quaternary_structure

Measuring ***gene expression*** (using mRNA) relies on sequence data analysis

Time (development)

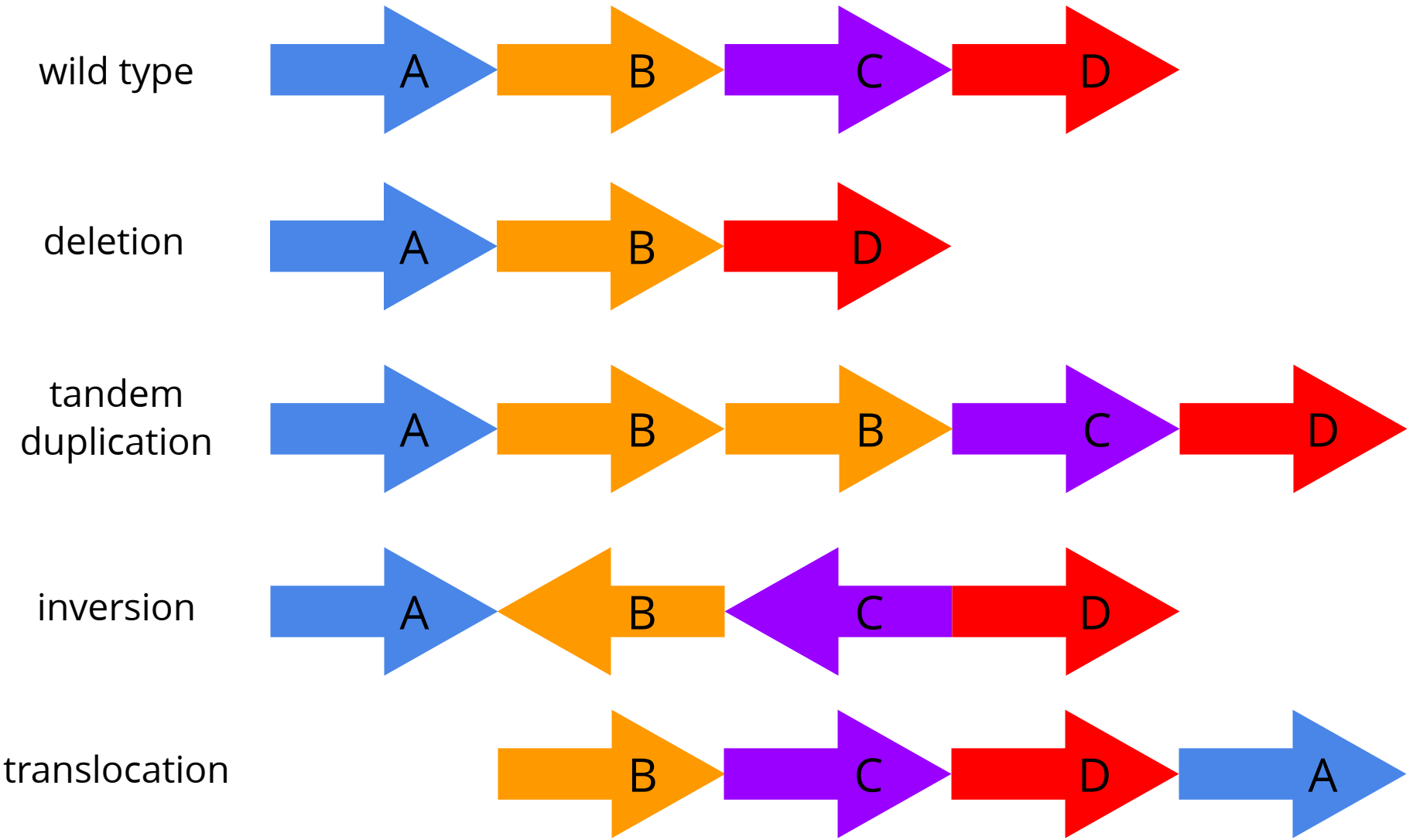


Local protein activity

No protein activity

Schiemann *et al.* (2017 *PNAS*)

Structural variation of genes among chromosomes studied by analyzing sequence data

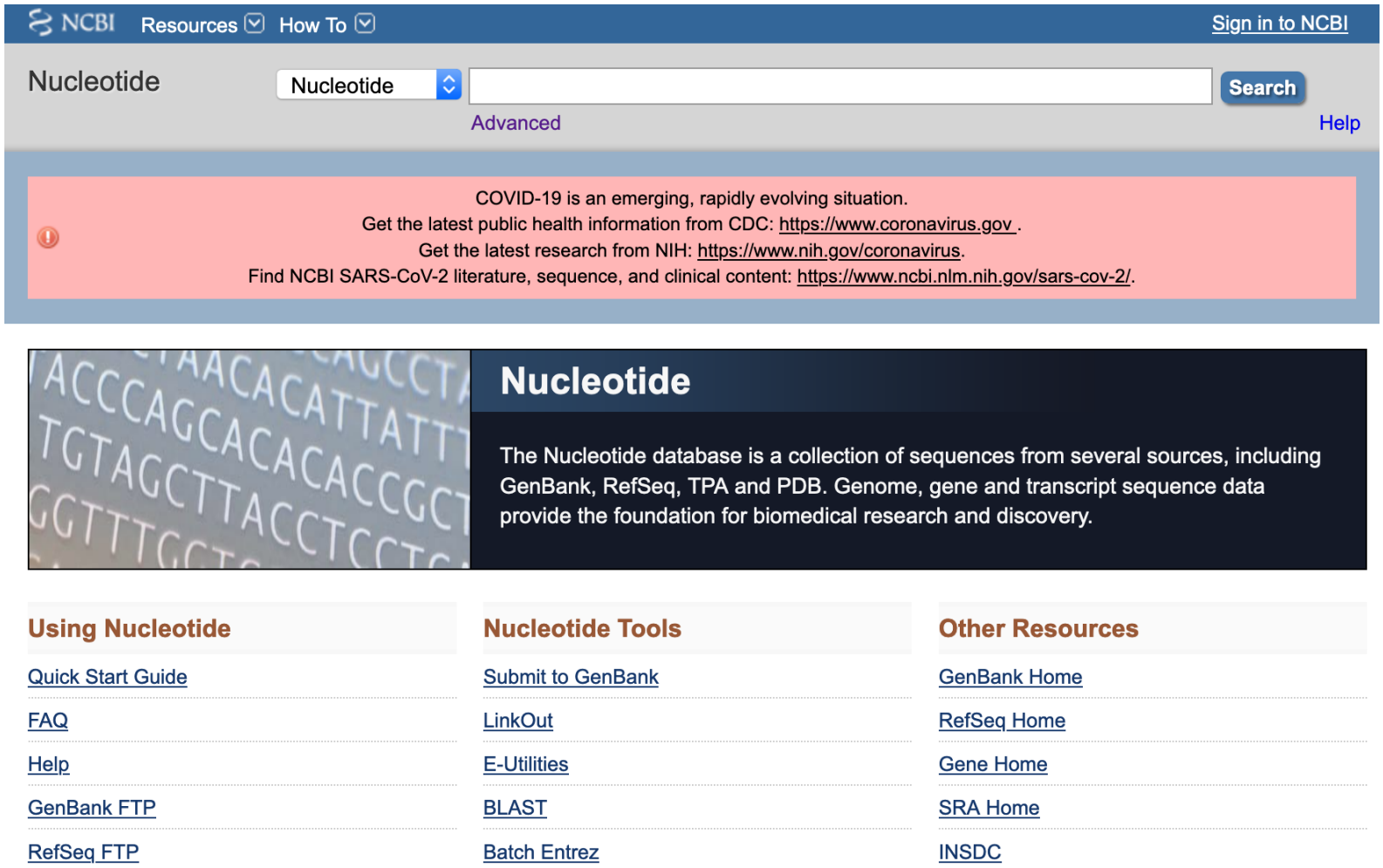


Sequence variation

```
                          *       * ***
Felis_cattus_cytB     tccgttattcattcaatc
Mus_musculus_cytB     tccgttatccacacaatc
Homo_sapiens_cytB     tccgttattctctcaatc
Bos_taurus_cytB       tctgttattcactcaatc
```

samples across species

NCBI GenBank



https://www.ncbi.nlm.nih.gov

GenBank sequences

GenBank publicly hosts annotated DNA sequences

- 218,642,238 sequences
- 654,057,069,549 bases

GenBank sequences are used extensively in research

- identifying anonymous sequences
- inferring gene function
- searching for drug targets
- expanding datasets
- meta-analyses

GenBank accession

Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds; mitochondrial

GenBank: JN034136.1

FASTA Graphics PopSet

Go to:

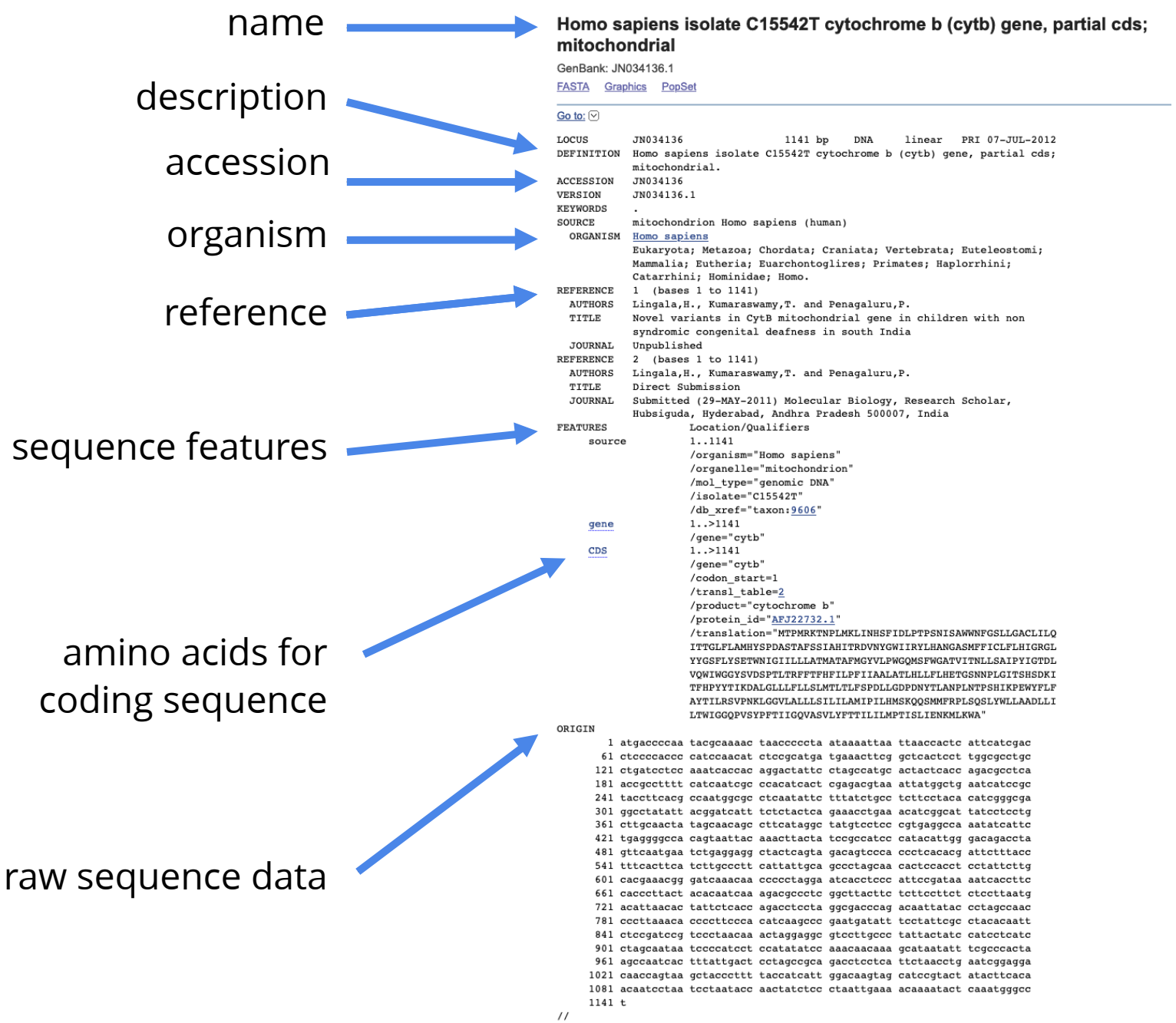
```
LOCUS       JN034136                1141 bp    DNA     linear   PRI 07-JUL-2012
DEFINITION  Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds;
            mitochondrial.
ACCESSION   JN034136
VERSION     JN034136.1
KEYWORDS    .
SOURCE      mitochondrion Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1141)
  AUTHORS   Lingala,H., Kumaraswamy,T. and Penagaluru,P.
  TITLE     Novel variants in CytB mitochondrial gene in children with non
            syndromic congenital deafness in south India
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1141)
  AUTHORS   Lingala,H., Kumaraswamy,T. and Penagaluru,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-MAY-2011) Molecular Biology, Research Scholar,
            Hubsiguda, Hyderabad, Andhra Pradesh 500007, India
```

GenBank accession (cont'd)

```
FEATURES             Location/Qualifiers
     source          1..1141
                     /organism="Homo sapiens"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /isolate="C15542T"
                     /db_xref="taxon:9606"
     gene            1..>1141
                     /gene="cytb"
     CDS             1..>1141
                     /gene="cytb"
                     /codon_start=1
                     /transl_table=2
                     /product="cytochrome b"
                     /protein_id="AFJ22732.1"
                     /translation="MTPMRKTNPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQ
                     ITTGLFLAMHYSPDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGL
                     YYGSFLYSETWNIGIILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDL
                     VQWIWGGYSVDSPTLTRFFTFHFILPFIIAALATLHLLFLHETGSNNPLGITSHSDKI
                     TFHPYYTIKDALGLLLFLLSLMTLTLFSPDLLGDPDNYTLANPLNTPSHIKPEWYFLF
                     AYTILRSVPNKLGGVLALLLSILILAMIPILHMSKQQSMMFRPLSQSLYWLLAADLLI
                     LTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLIENKMLKWA"
```

GenBank accession (cont'd)

```
ORIGIN
        1 atgaccccaa tacgcaaaac taacccccta ataaaattaa ttaaccactc attcatcgac
       61 ctccccaccc catccaacat ctccgcatga tgaaacttcg gctcactcct tggcgcctgc
      121 ctgatcctcc aaatcaccac aggactattc ctagccatgc actactcacc agacgcctca
      181 accgcctttt catcaatcgc ccacatcact cgagacgtaa attatggctg aatcatccgc
      241 taccttcacg ccaatggcgc ctcaatattc tttatctgcc tcttcctaca catcgggcga
      301 ggcctatatt acggatcatt tctctactca gaaacctgaa acatcggcat tatcctcctg
      361 cttgcaacta tagcaacagc cttcataggc tatgtcctcc cgtgaggcca aatatcattc
      421 tgaggggcca cagtaattac aaacttacta tccgccatcc catacattgg gacagaccta
      481 gttcaatgaa tctgaggagg ctactcagta gacagtccca ccctcacacg attctttacc
      541 tttcacttca tcttgccctt cattattgca gccctagcaa cactccacct cctattcttg
      601 cacgaaacgg gatcaaacaa ccccctagga atcacctccc attccgataa aatcaccttc
      661 cacccttact acacaatcaa agacgccctc ggcttacttc tcttccttct ctccttaatg
      721 acattaacac tattctcacc agacctccta ggcgacccag acaattatac cctagccaac
      781 cccttaaaca ccccttccca catcaagccc gaatgatatt tcctattcgc ctacacaatt
      841 ctccgatccg tccctaacaa actaggaggc gtccttgccc tattactatc catcctcatc
      901 ctagcaataa tccccatcct ccatatatcc aaacaacaaa gcataatatt tcgcccacta
      961 agccaatcac tttattgact cctagccgca gacctcctca ttctaacctg aatcggagga
     1021 caaccagtaa gctacccttt taccatcatt ggacaagtag catccgtact atacttcaca
     1081 acaatcctaa tcctaatacc aactatctcc ctaattgaaa acaaaatact caaatgggcc
     1141 t
//
```


name
description
accession
organism
reference
sequence features
amino acids for coding sequence
raw sequence data
Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds; mitochondrial
GenBank: JN034136.1
FASTA Graphics PopSet
Go to:
LOCUS JN034136 1141 bp DNA linear PRI 07-JUL-2012
DEFINITION Homo sapiens isolate C15542T cytochrome b (cytb) gene, partial cds; mitochondrial.
ACCESSION JN034136
VERSION JN034136.1
KEYWORDS .
SOURCE mitochondrion Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1141)
AUTHORS Lingala,H., Kumaraswamy,T. and Penagaluru,P.
TITLE Novel variants in CytB mitochondrial gene in children with non syndromic congenital deafness in south India
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 1141)
AUTHORS Lingala,H., Kumaraswamy,T. and Penagaluru,P.
TITLE Direct Submission
JOURNAL Submitted (29-MAY-2011) Molecular Biology, Research Scholar, Hubsiguda, Hyderabad, Andhra Pradesh 500007, India
FEATURES Location/Qualifiers
source 1..1141
/organism="Homo sapiens"
/organelle="mitochondrion"
/mol_type="genomic DNA"
/isolate="C15542T"
/db_xref="taxon:9606"
gene 1..>1141
/gene="cytb"
CDS 1..>1141
/gene="cytb"
/codon_start=1
/transl_table=2
/product="cytochrome b"
/protein_id="AFJ22732.1"
/translation="MTPMRKTNPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQ ITTGLFLAMHYSPDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGL YYGSFLYSETWNIGIILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDL VQWIWGGYSVDSPTLTRFFTFHFILPFIIAALATLHLLFLHETGSNNPLGITSHSDKI TFHPYYTIKDALGLLLFLLSLMTLTLFSPDLLGDPDNYTLANPLNTPSHIKPEWYFLF AYTILRSVPNKLGGVLALLLSILILAMIPILHMSKQQSMMFRPLSQSLYWLLAADLLI LTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLIENKMLKWA"
ORIGIN
1 atgaccccaa tacgcaaaac taacccccta ataaaattaa ttaaccactc attcatcgac
61 ctccccaccc catccaacat ctccgcatga tgaaacttcg gctcactcct tggcgcctgc
121 ctgatcctcc aaatcaccac aggactattc ctagccatgc actactcacc agacgcctca
181 accgcctttt catcaatcgc ccacatcact cgagacgtaa attatggctg aatcatccgc
241 taccttcacg ccaatggcgc ctcaatattc tttatctgcc tcttcctaca catcgggcga
301 ggcctatatt acggatcatt tctctactca gaaacctgaa acatcggcat tatcctcctg
361 cttgcaacta tagcaacagc cttcataggc tatgtcctcc cgtgaggcca aatatcattc
421 tgaggggcca cagtaattac aaacttacta tccgccatcc catacattgg gacagaccta
481 gttcaatgaa tctgaggagg ctactcagta gacagtccca ccctcacacg attctttacc
541 tttcacttca tcttgccctt cattattgca gccctagcaa cactccacct cctattcttg
601 cacgaaacgg gatcaaacaa ccccctagga atcacctccc attccgataa aatcaccttc
661 cacccttact acacaatcaa agacgccctc ggcttacttc tcttccttct ctccttaatg
721 acattaacac tattctcacc agacctccta ggcgacccag acaattatac cctagccaac
781 cccttaaaca ccccttccca catcaagccc gaatgatatt tcctattcgc ctacacaatt
841 ctccgatccg tccctaacaa actaggaggc gtccttgccc tattactatc catcctcatc
901 ctagcaataa tccccatcct ccatatatcc aaacaacaaa gcataatatt tcgcccacta
961 agccaatcac tttattgact cctagccgca gacctcctca ttctaacctg aatcggagga
1021 caaccagtaa gctacccttt taccatcatt ggacaagtag catccgtact atacttcaca
1081 acaatcctaa tcctaatacc aactatctcc ctaattgaaa acaaaatact caaatgggcc
1141 t
//

NCBI collections

Viruses > Riboviria > Orthornavirae > Kitrinoviricota > Alsuviricetes > Martellivirales > Togaviridae >

Alphavirus - 33 complete genomes

Retrieve sequences: -- Select data set from the list --

Species [32] unclassified Alphavirus [1]

* The list view for each taxonomy node shows only the next level of sublineages.
* Unclassified/unassigned names are written in copper

Sort the genomes list by: taxonomy | Filter by host: all hosts | >Download

Show / hide all segment lists

Genome	Accession	RefSeq type	Source information	Segm	Length	Protein	Neighbors	Host	Created	Updated
Aura virus	NC_003900	complete		-	11824 nt	3	1		02/09/1999	08/13/2018
Barmah Forest virus	NC_001786	complete	strain:BH2193	-	11488 nt	4	32	human, invertebrates, vertebrates	01/14/1997	08/13/2018
Bebaru virus	NC_016962	complete		-	11877 nt	3	-		03/09/2012	08/13/2018
Cabassou virus	NC_038670	complete	strain:CaAr 508	-	11385 nt	2	-		08/24/2018	08/24/2018
Chikungunya virus	NC_004162	complete	strain:S27-African prototype	-	11826 nt	2	876	human, invertebrates, vertebrates	09/06/2002	10/29/2018
Eastern equine encephalitis virus	NC_003899	complete	strain:ssp. North American variant	-	11675 nt	4	453	human, invertebrates, vertebrates	12/16/1991	08/13/2018
Eilat virus	NC_018615	complete	isolate:EO329	-	11634 nt	3	1	invertebrates	09/20/2012	08/13/2018
Everglades virus	NC_038671	complete	strain:Everglades Fe3-7c	-	11395 nt	2	-		08/24/2018	08/24/2018
Fort Morgan virus	NC_013528	complete	isolate:CM4-146	-	11381 nt	3	1	invertebrates, vertebrates	11/25/2009	08/13/2018
Getah virus	NC_006558	complete	isolate:swine	-	11597 nt	3	36	invertebrates, vertebrates	12/17/2004	08/13/2018
Highlands J virus	NC_012561	complete	isolate:585-01	-	11526 nt	3	9	invertebrates, vertebrates	04/15/2009	08/13/2018
Madariaga virus	NC_023812	complete	strain:MADV/Cebus apella/BRA/BEAN5122/1956	-	11624 nt	2	29	human, invertebrates, vertebrates	03/20/2014	08/13/2018
Mayaro virus	NC_003417	complete		-	11411 nt	4	40	human, invertebrates, vertebrates	02/22/2002	08/13/2018
Middelburg virus	NC_024887	complete	isolate:ArB-8422	-	11550 nt	2	3	invertebrates, vertebrates	09/17/2014	08/13/2018
Mosso das Pedras virus	NC_038857	complete	strain:78V-3531	-	11465 nt	2	-		08/24/2018	08/24/2018
Mucambo virus	NC_038672	complete	strain:Mucambo BeAn 8	-	11391 nt	2	1	invertebrates	08/24/2018	08/24/2018
Ndumu virus	NC_016959	complete		-	11688 nt	4	-	invertebrates, vertebrates	03/09/2012	08/13/2018
Onyong-nyong virus	NC_001512	complete		-	11835 nt	3	3	human, invertebrates	08/02/1993	08/13/2018
Pixuna virus	NC_038673	complete	strain:Pixuna BeAr 35645	-	11344 nt	2	-	vertebrates	08/24/2018	08/24/2018
Rio Negro virus	NC_038674	complete	strain:AG80-663	-	11494 nt	2	-	invertebrates,	08/24/2018	08/24/2018

Entrez Direct: E-utilities on the Unix Command Line

Jonathan Kans, PhD[1]

Created: April 23, 2013; Updated: September 14, 2020.

Getting Started

Introduction

Entrez Direct (EDirect) provides access to the NCBI's suite of interconnected databases (publication, sequence, structure, gene, variation, expression, etc.) from a Unix terminal window. Search terms are entered as command-line arguments. Individual operations are connected with Unix pipes to allow construction of multi-step queries. Selected records can then be retrieved in a variety of formats.

EDirect also includes an argument-driven function that simplifies the extraction of data from document summaries or other results that are returned in structured XML format. This can eliminate the need for writing custom software to answer ad hoc questions. Queries can move seamlessly between EDirect commands and Unix utilities or scripts to perform actions that cannot be accomplished entirely within Entrez.

www.ncbi.nlm.nih.gov/books/NBK179288/

wget, web-get

Downloads files over HTTP/S and FTP;

simple interface, allows for recursive copy

```
# download a file located at the target URL
$ wget https://website.com/download_me.txt
# save the file to a new location (`-O`)
$ wget —O my_downloaded_file.txt https://website.com/download_me.txt
# save all files listed in target file
$ wget —i MultipleDownloads.txt
# recursively download all files and subdirectories located
# in the target directory (`-r -np`)
$ wget -r -np  http://example.com/configs/.vim/
# ...same as above, but ignore some files (`-R`)
$ wget -r -np -R "index.html*" http://example.com/configs/.vim/
```

curl, transfer a URL

Downloads files over many ports;

many features, including resume download

download interrupted

```
# download file from web address
$ curl -O https://releases.ubuntu.com/20.04.1/ubuntu-20.04.1-desktop-amd64.iso
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  3 2656M    3 89.8M    0     0  11.2M      0  0:03:56  0:00:08  0:03:48 16.9M^C
# resume download using `-C -`
$ curl -C - -O https://releases.ubuntu.com/20.04.1/ubuntu-20.04.1-desktop-amd64.iso
** Resuming transfer from byte position 104992768
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
 10 2555M   10  255M    0     0  24.0M      0  0:01:46  0:00:10  0:01:36 27.4M
# file transfer FTP
$ curl -u demo:password -O ftp://test.rebex.net/readme.txt
```

Lab 4A

github.com/WUSTL-Biol4220/home/labs/lab_03A.md