BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
SPRING 2020
SESSION 2

THE UNIVERSITY OF CHICAGO

© T.A. BINKOWSKI, 2020

SESSION 2

GENOMES, SEQUENCES AND DATABASES

- Objectives
  - Introduce sequencing technologies
  - Learn about the most important DNA databases
  - Introduce import nomenclature
    - Accession numbers, versions and gi
  - Understand the important of RefSeq project
  - Naviage accession numbers for any gene (or protein) from any organism via Entrez

GENOMES, SEQUENCES AND DATABASES

- There is ambiguity in many terms and nomenclature

BIOLOGICAL DATABASES

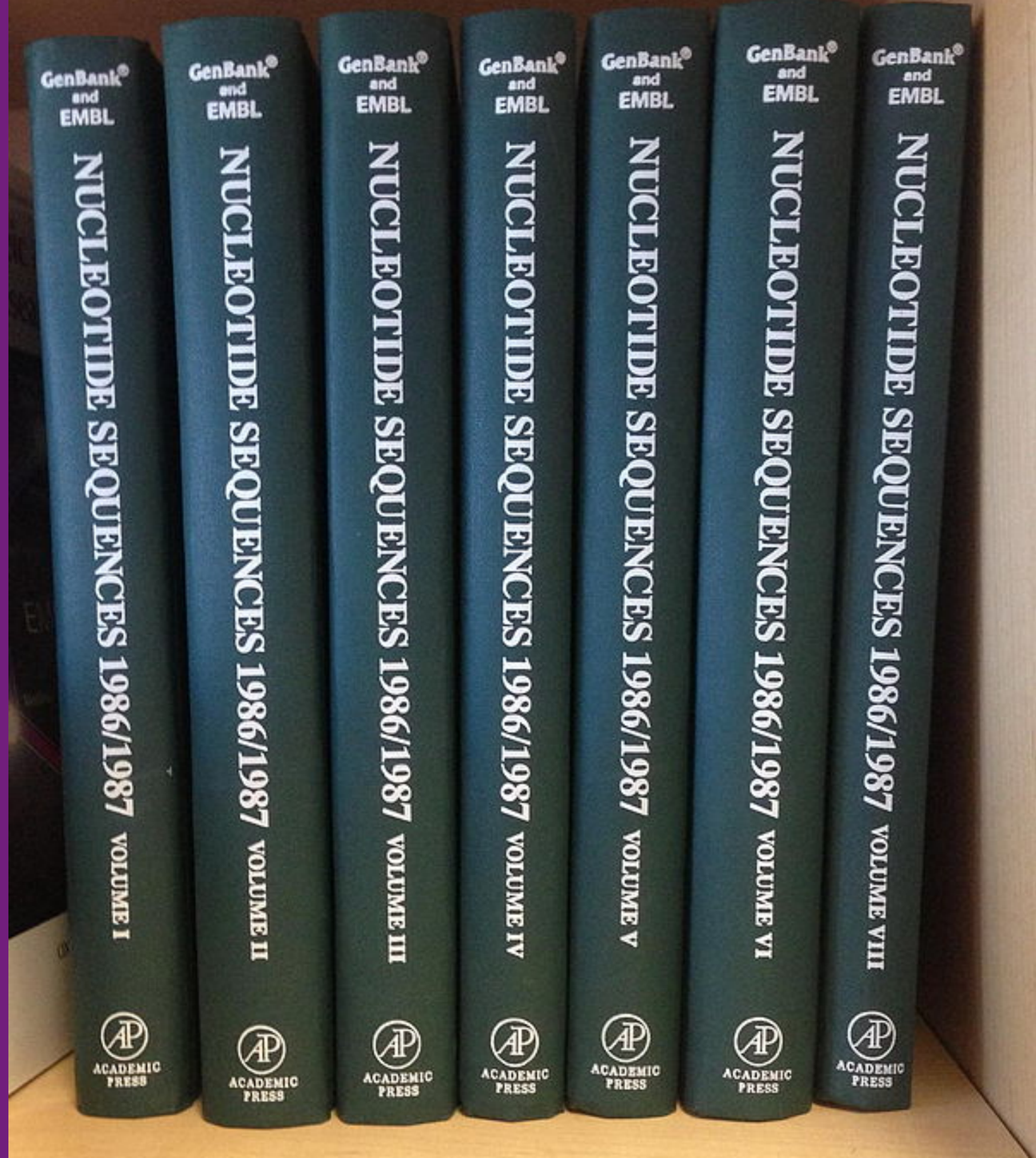
BIOLOGICAL DATABASES

- All of that technology ends up giving us lots and lots of strings

```
>gi|568815587:c5227071-5225466 Homo sapiens chromosome 11, GRCh38 Primary Assembly
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA
TCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAAT
ATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAAT
CATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAA
TTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATA
CTTTCCCTAATCTCTTTCTTTCAGGGCAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAG
AATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAATATTTCTGCATATAAAT
TGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTT
ATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTT
ATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC
```

BIOLOGICAL DATABASES

- Genbank and EMBL: NucleotideSequences 1986/1987 Volumes I to VII
- The GenBank release notes for release 162.0 (October 2007)
  - "from 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months".
- As of February 2018, GenBank release 200.0
  - 253,630,708,098 bases
  - 207,040,555 sequences



BIOLOGICAL DATABASES

- This won't fit in an excel spreadsheet
  - Probably a flat file

Top organisms in GenBank (Release 191)[8]

Organism	base pairs
Homo sapiens	16,310,774,187
Mus musculus	9,974,977,889
Rattus norvegicus	6,521,253,272
Bos taurus	5,386,258,455
Zea mays	5,062,731,057
Sus scrofa	4,887,861,860
Danio rerio	3,120,857,462
Strongylocentrotus purpuratus	1,435,236,534
Macaca mulatta	1,256,203,101
Oryza sativa Japonica Group	1,255,686,573
Nicotiana tabacum	1,197,357,811

BIOLOGICAL DATABASES

WHY SO MANY DATABASES?

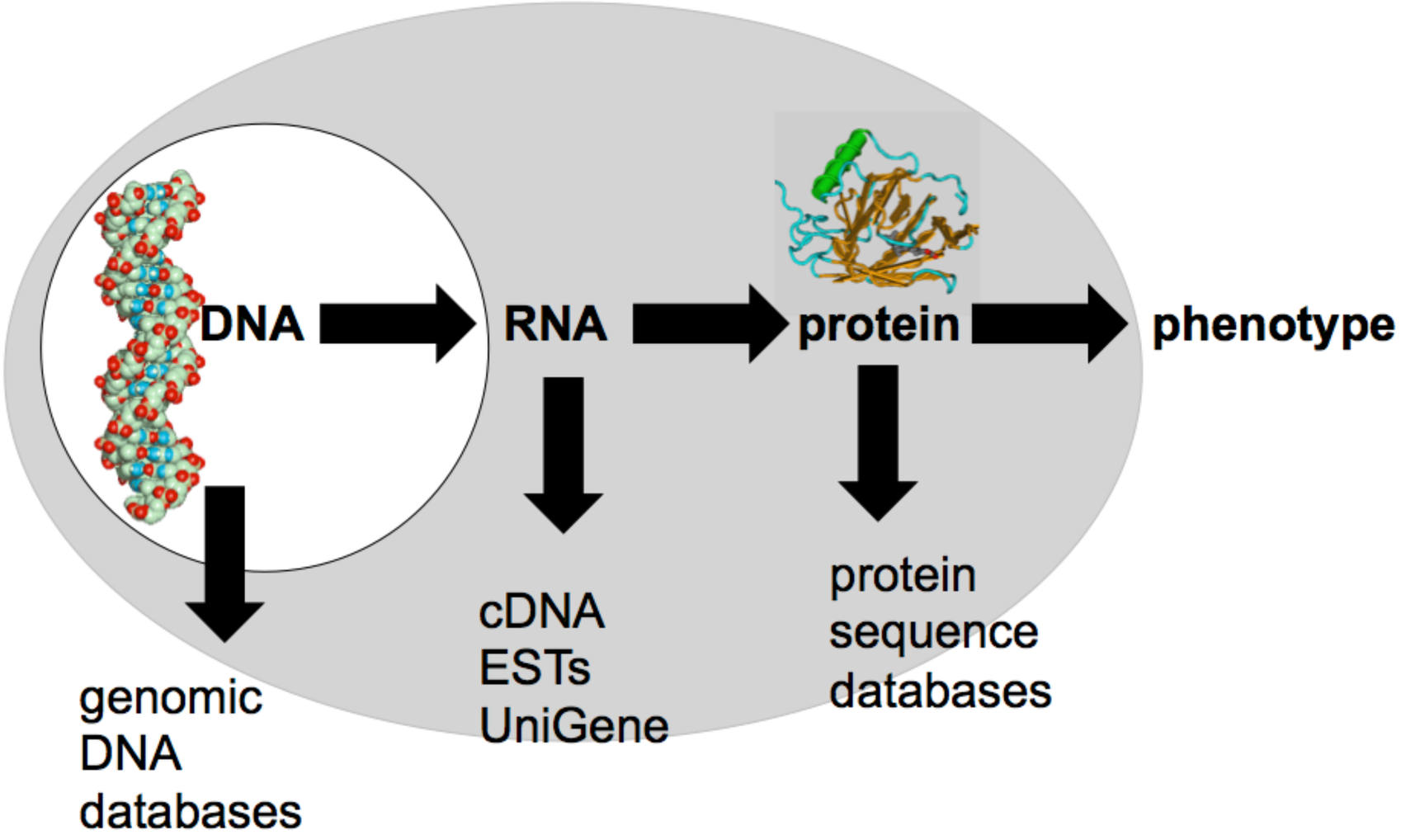
- Genome and genomic sequences
- Gene sequences, mutations
- Gene regulation
- Gene expression (where and when) Intron splice variants
- Protein sequence, post-translational modifications

OVERLAPPING AND RELATED DATA

- Protein tertiary structure (3D) Protein networks
- Protein localization
- Enzyme Kinetics
- Metabolites, metabolic networks Diseases
- Literature

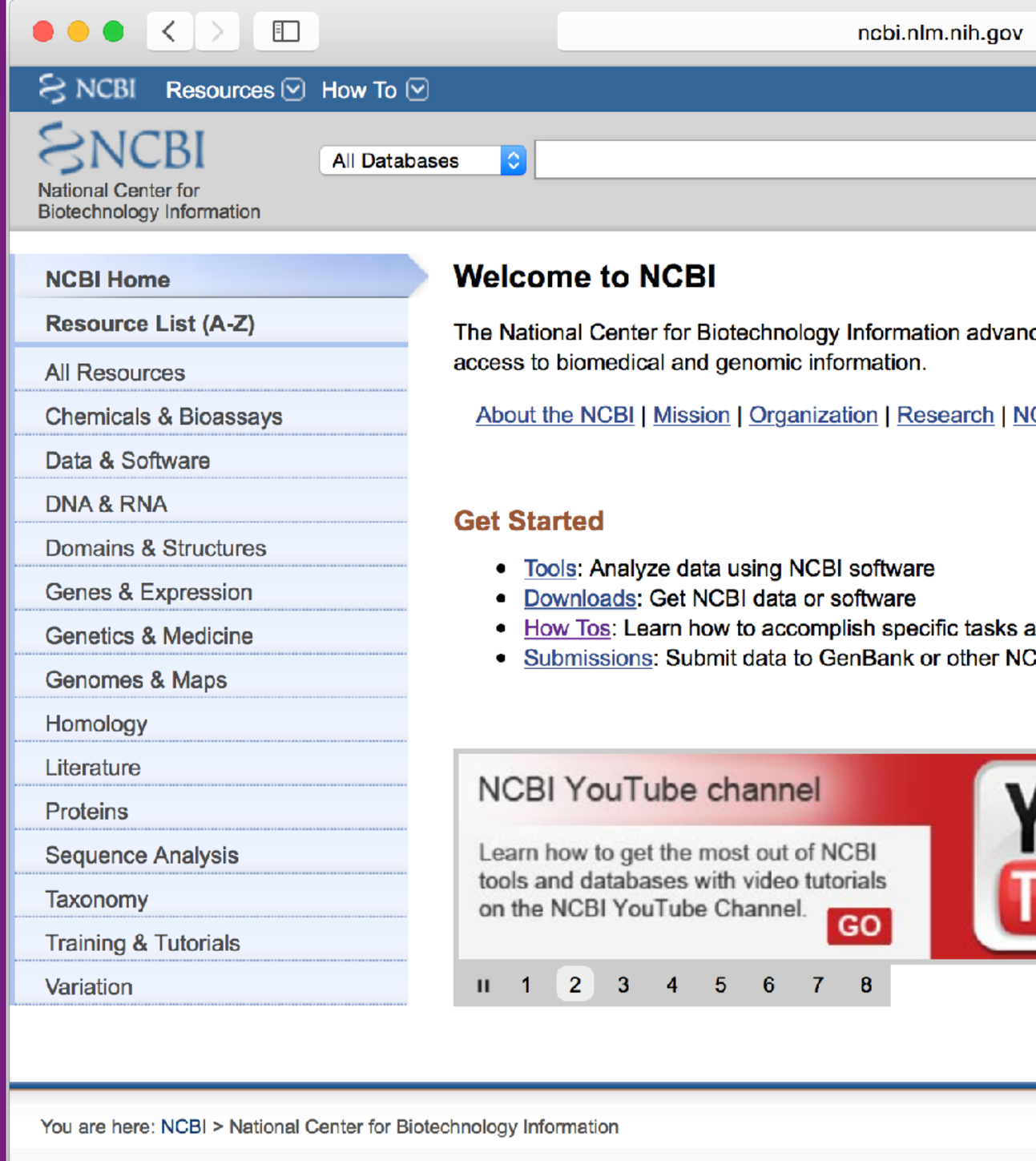
BIOLOGICAL DATABASES

- Hierarchy of data at different biological levels and time scales



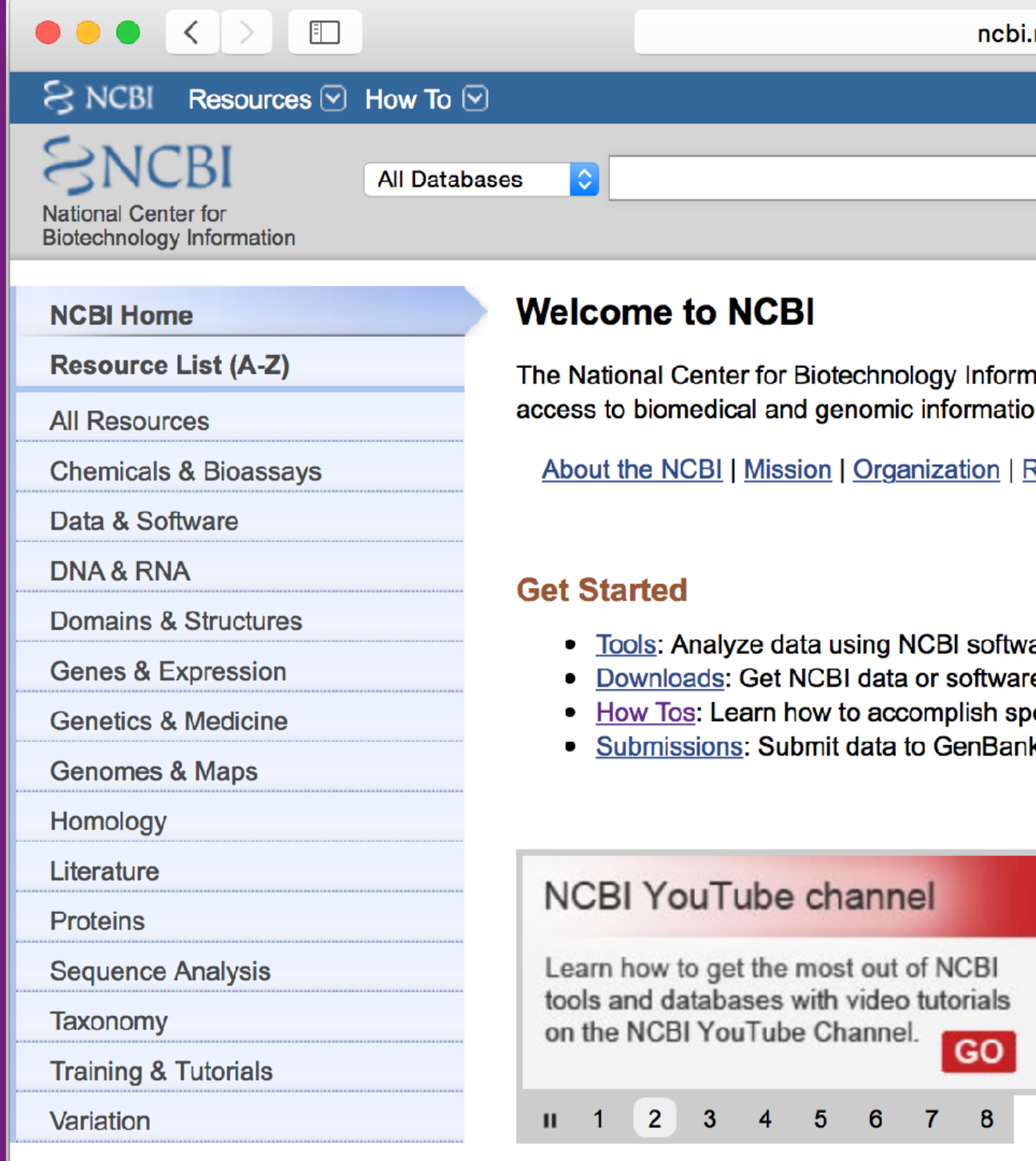
BIOLOGICAL DATABASES

- Three major international DNA databases
  - EMBL (EBI)
  - GenBank (NCBI)
  - DDBJ (Japan)
- Coordinated effort to share and normalize data
  - Official releases
  - Daily data transfers/updates



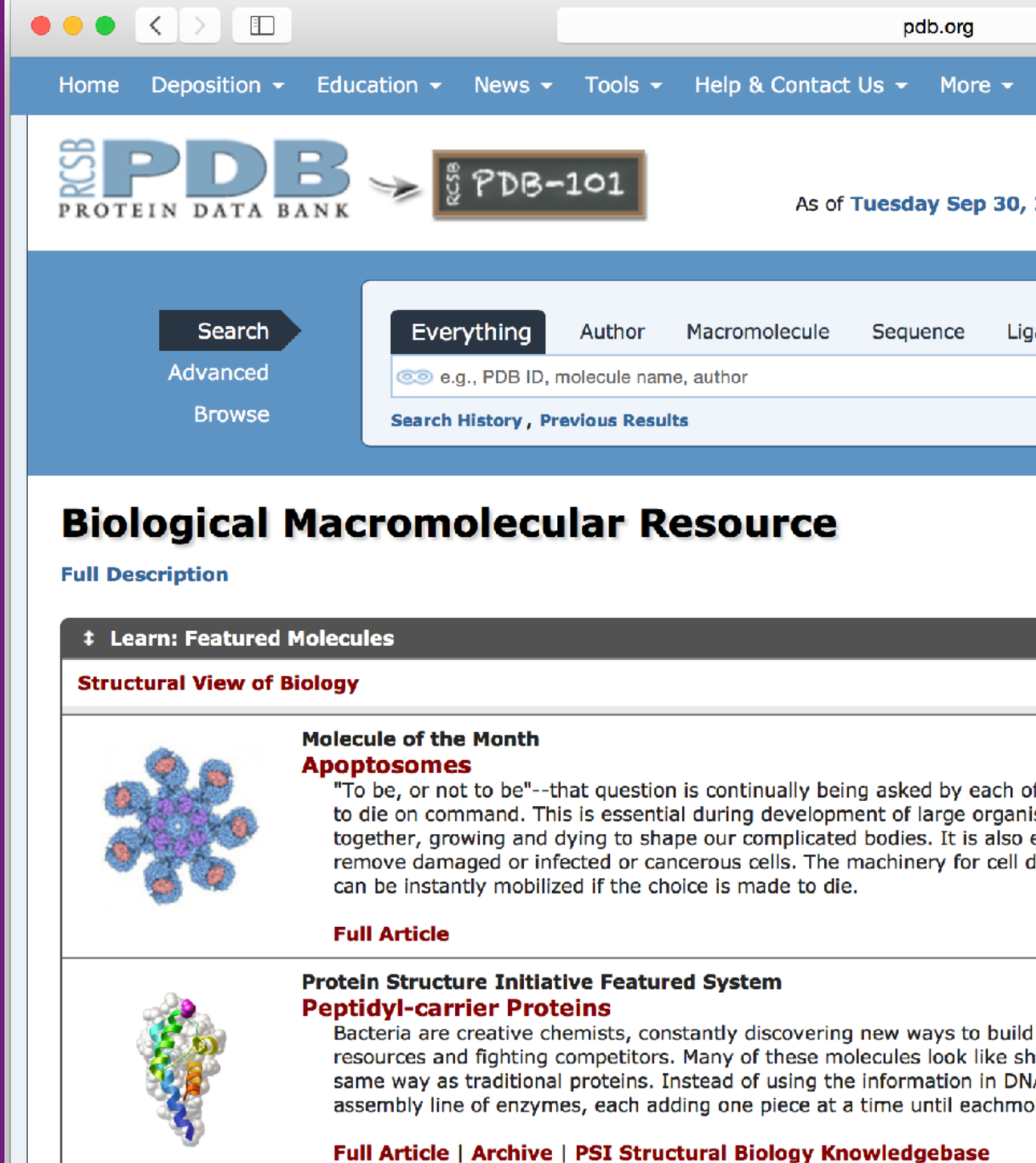
BIOLOGICAL DATABASES

- Data sources for DNA databases
  - Direct submission from researchers
  - Genome sequencing labs
  - Scientific literature
  - Patent applications



BIOLOGICAL DATABASES

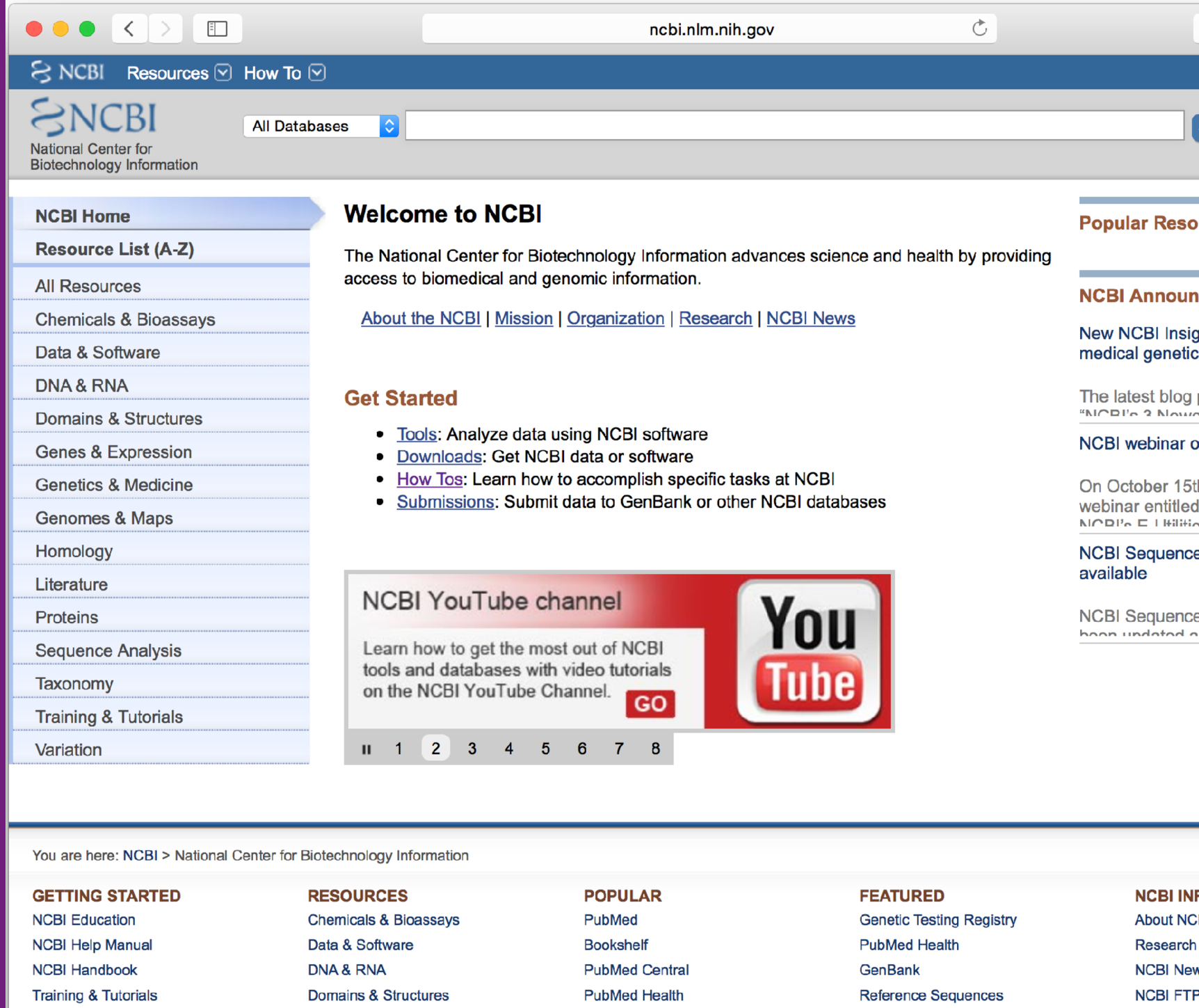
- Many specialized databases
  - Proteins - SwissProt, PDB
  - Diseases
  - Organism
  - …
- Literature database
  - PubMed, GoogleScholar, …



NCBI

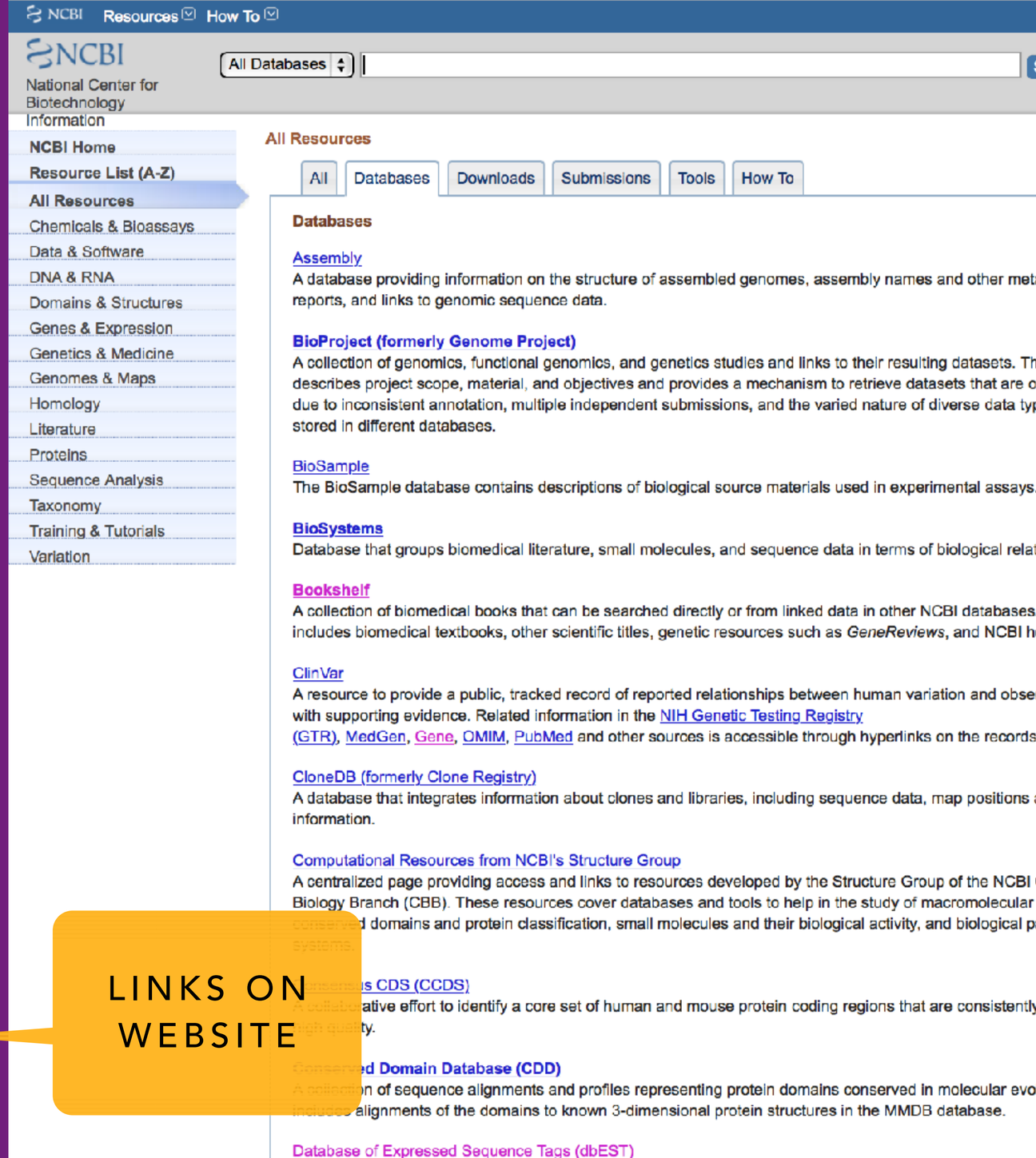
NCBI

- The National Center for Biotechnology Information (NCBI)
  - Advances science and health by providing access to biomedical and genomic information



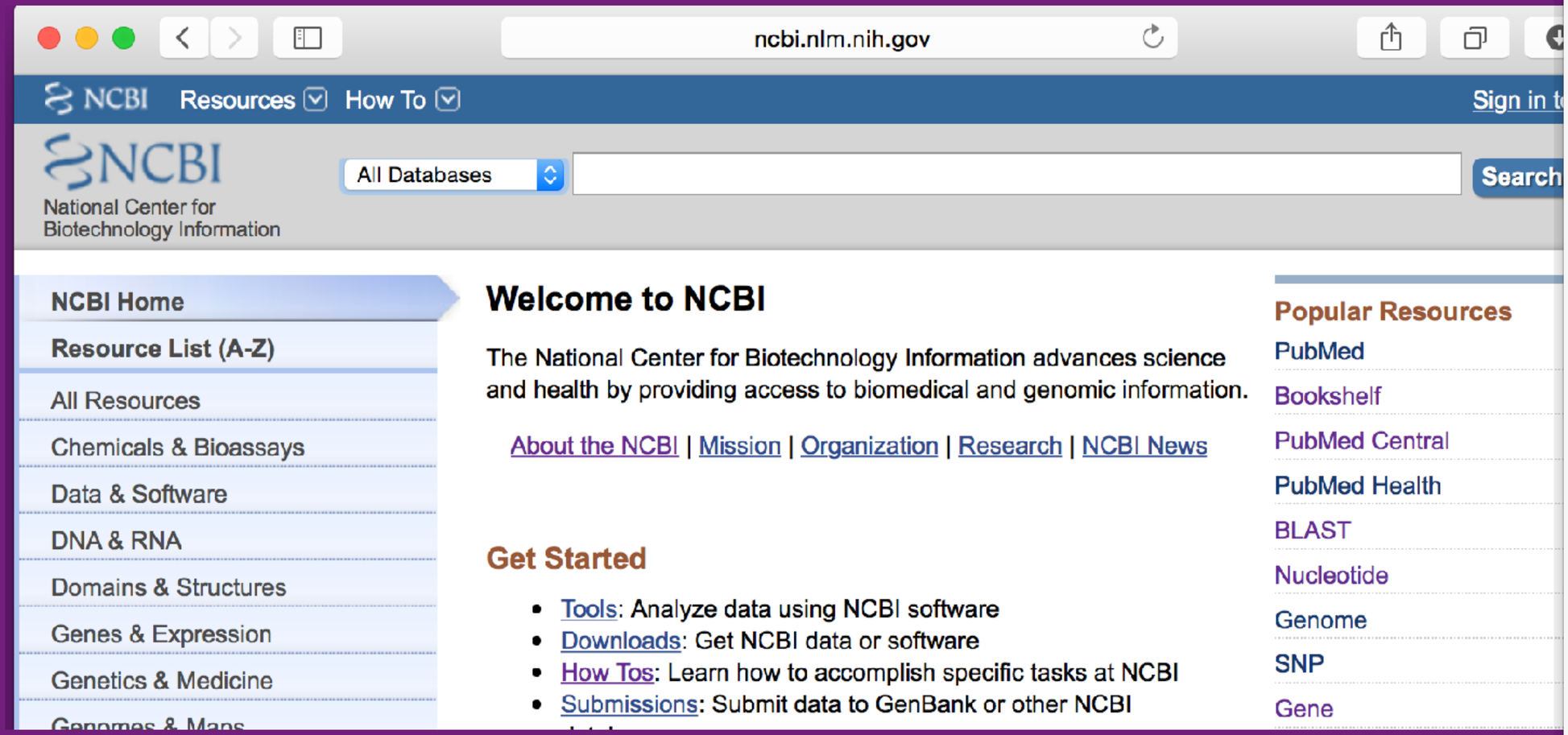
NCBI

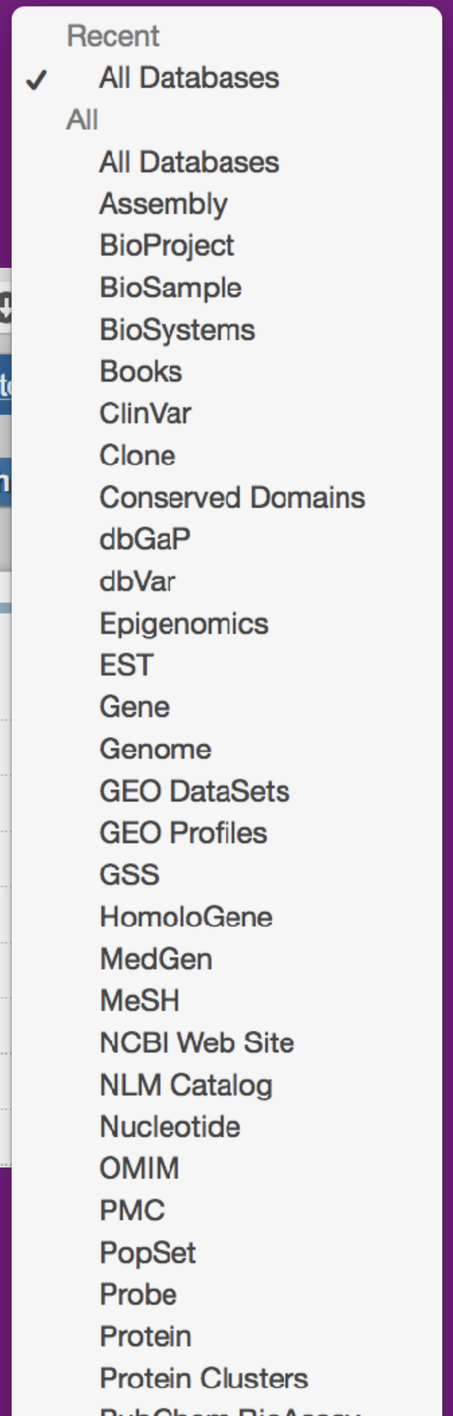
- Resources
  - Databases
  - Computational Tools
  - Books
  - Literature
- NCBI Handbook
  - http://www.ncbi.nlm.nih.gov/books/NBK21106/
- Tutorials
  - http://www.ncbi.nlm.nih.gov/books/NBK45612/



LINKS ON WEBSITE

NCBI





- Databases are subdivided for easier searching

NCBI

ORGANISM

- PRI - primate (human,monkey)
- ROD - rodent (mouse,rat)
- MAM - other mammalian (bovine,cat)
- VRT - other vertebrate (chicken)
- INV - invertebrate
- PLN - plant, fungal, and alga
- BCT - bacteria
- VRL - viruses
- PHG - bacteriophage
- SYN - synthetic (plasmids, vectors)
- UNA - unannotated sequences
- PAT - patent sequences

- EST - Expressed Sequence Tags
- STS - Sequence Tagged Sites
- GSS - Genome Survey Sequences
- HTG - High Throughput Genomic Sequences
- HTC - High Throughput cDNA Sequences

TYPES OF DATA

- Databases are subdivided for easier searching

NCBI

MOST USED RESOURCES FOR THIS CLASS

ncbi.nlm.nih.gov

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

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 | Research | NCBI News

Get Started

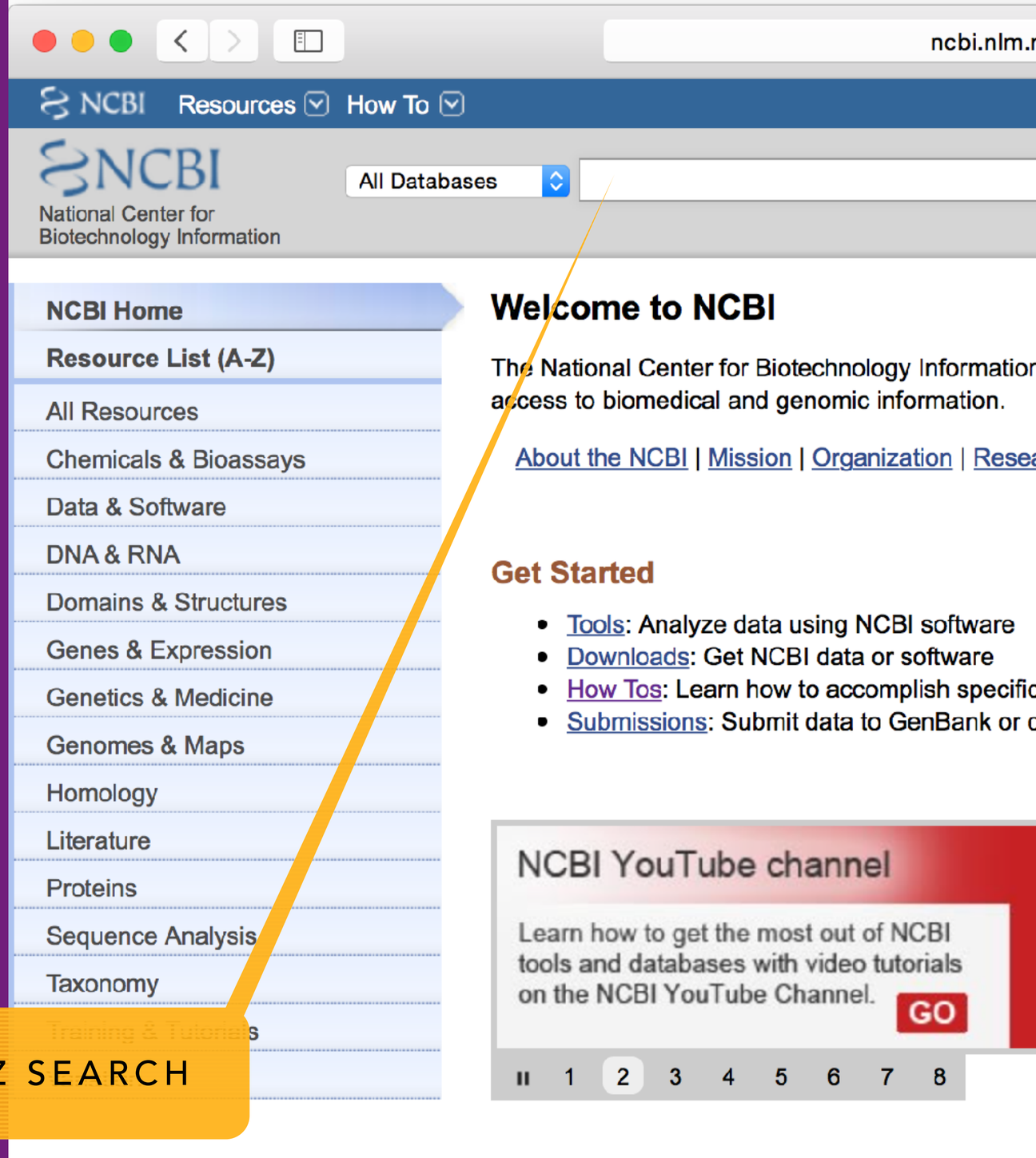
- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene

NCBI

- Entrez search and retrieval system integrates
  - DNA, protein sequence databases
  - Protein structural data
  - Genome assemblies
  - Literature
  - …
  - everything

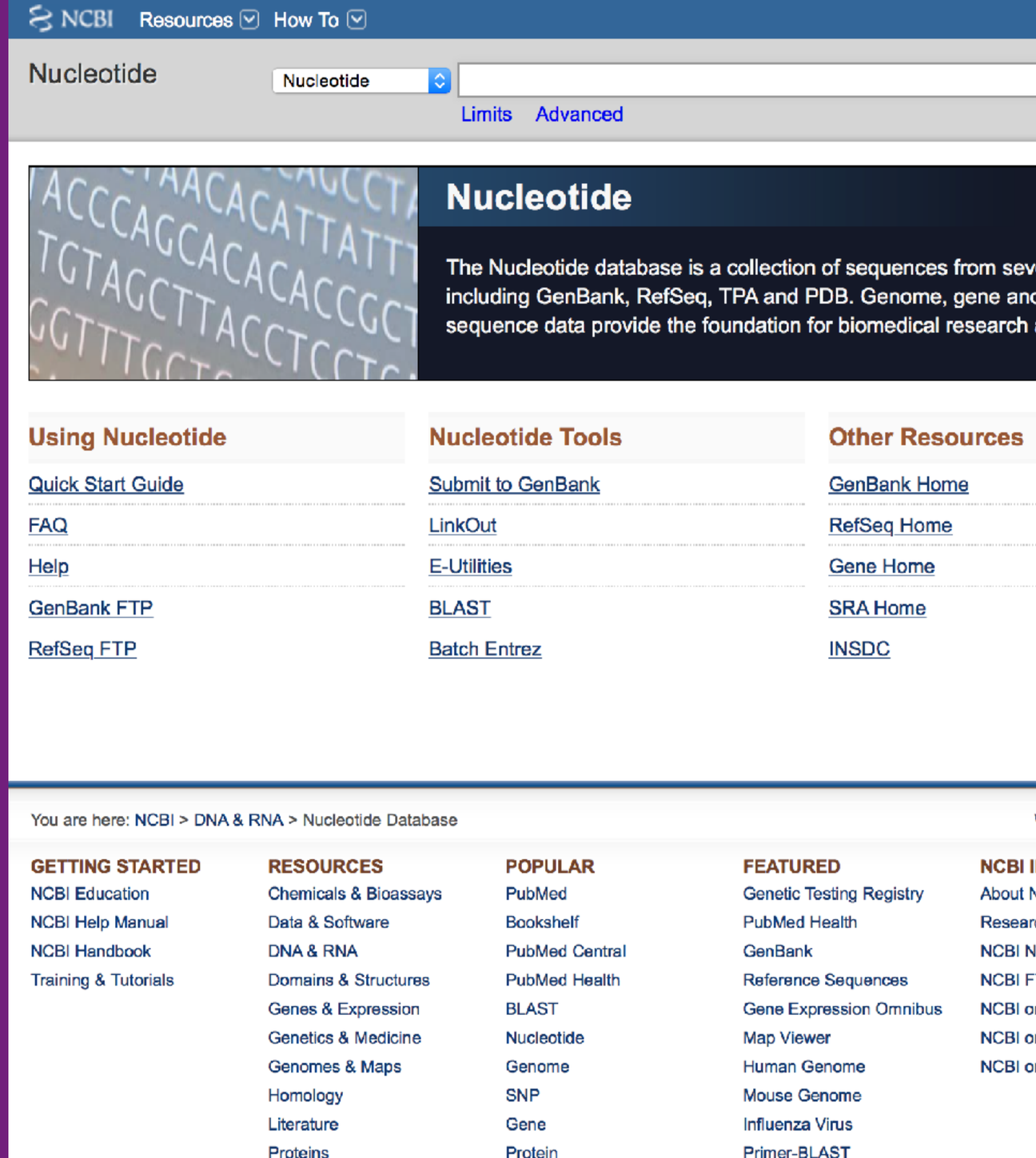


ENTREZ SEARCH

GenBank

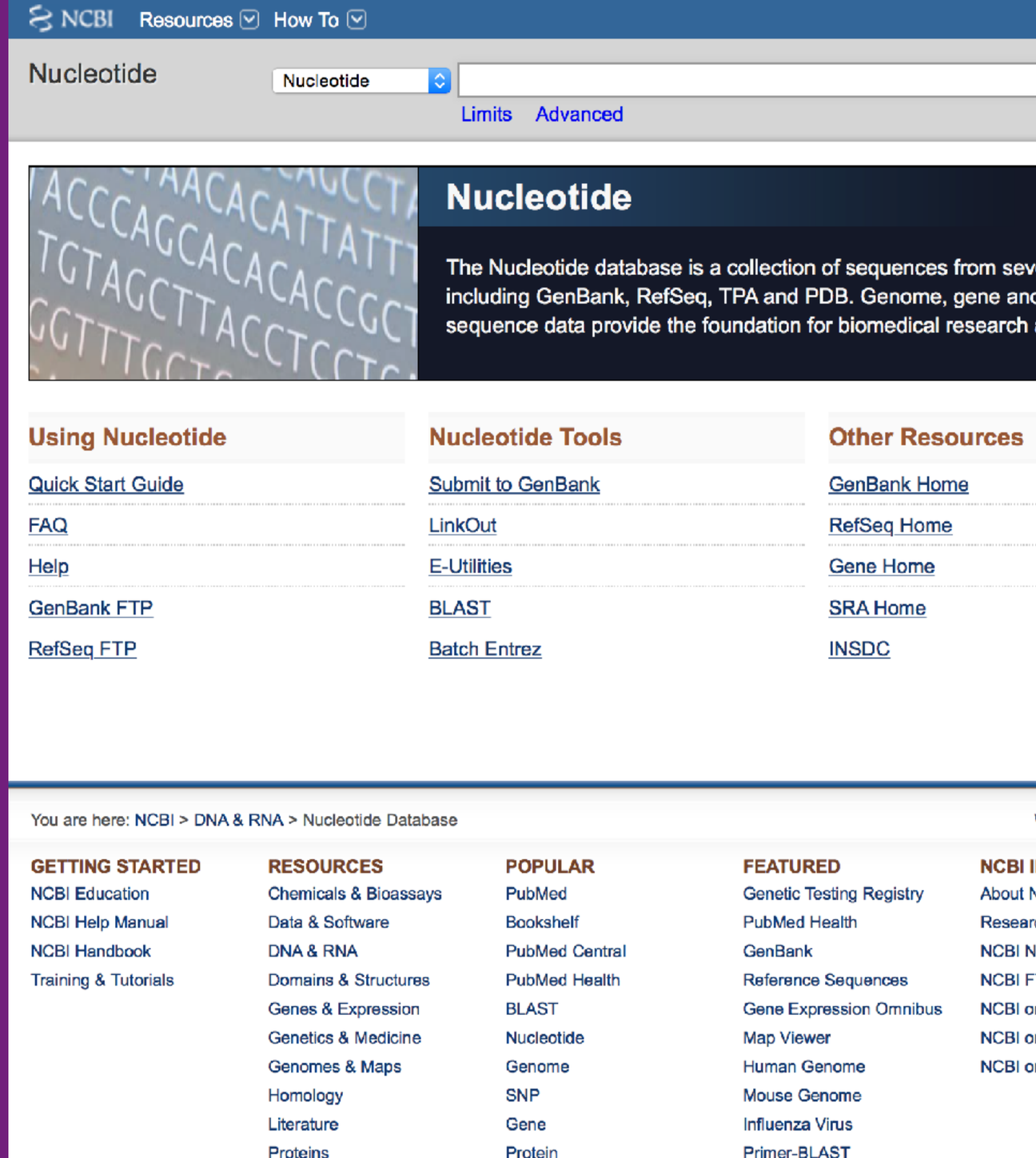
GENBANK

- GenBank is the NIH genetic sequence database
  - Annotated collection of all publicly available DNA sequences

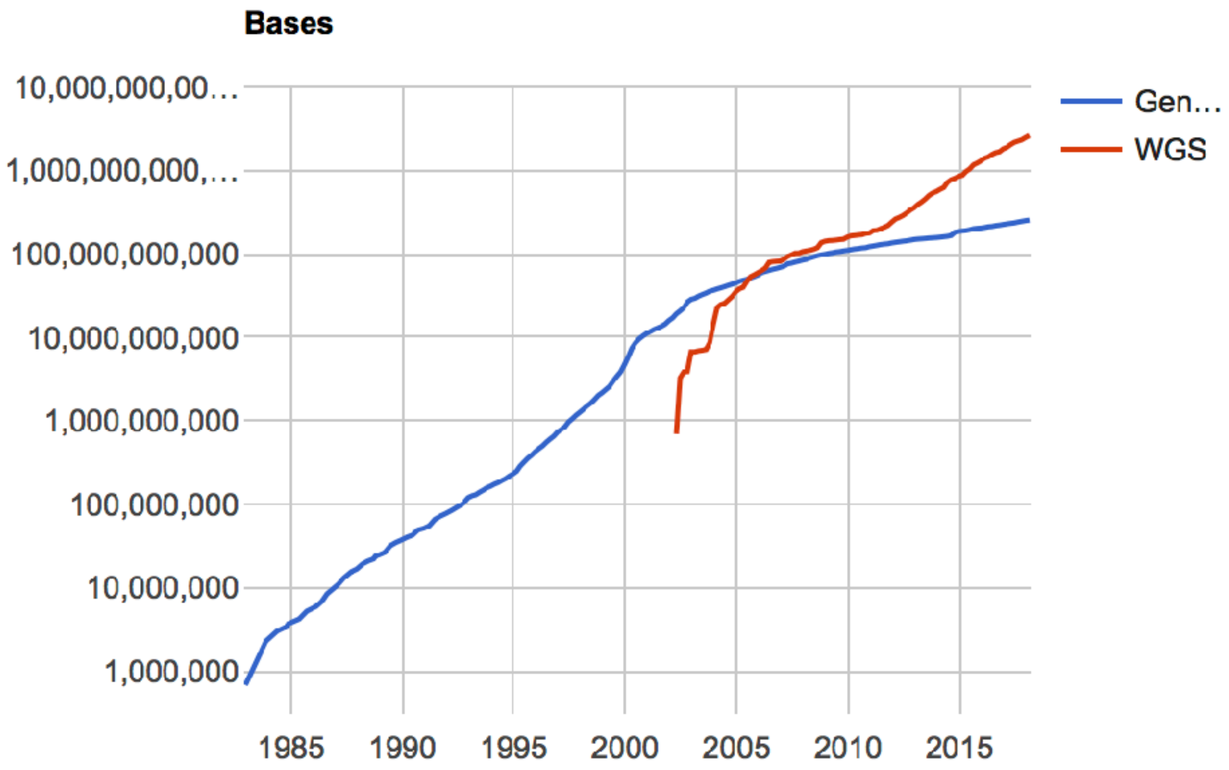


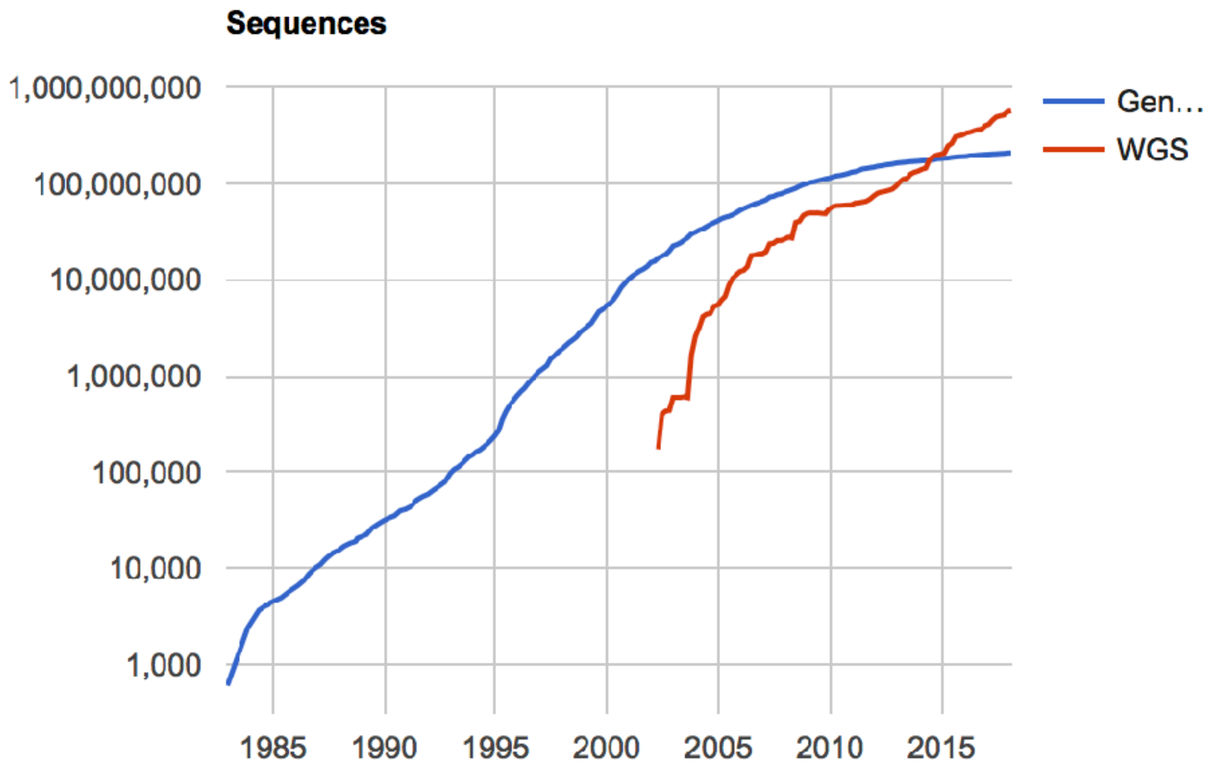
GENBANK

- GenBank is part of the International Nucleotide Sequence Database Collaboration (INSDC)
  - DNA DataBank of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL)
  - GenBank at NCBI
  - Exchange data on a daily basis



GENBANK


Bases
10,000,000,00…
1,000,000,000,…
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
1985
1990
1995
2000
2005
2010
2015
Gen…
WGS


Sequences
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
1985
1990
1995
2000
2005
2010
2015
Gen…
WGS

GENBANK

- GenBank records store in GenBank Flat File Format
  - Contains sequence and information about the sequence
- Example: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

```
LOCUS       SCU49845     5028 bp    DNA             PLN
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial c
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Sa
            Saccharomycetales; Saccharomycetaceae; Saccharomyc
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C
  TITLE     Cloning and sequence of REV7, a gene whose functio
            DNA damage-induced mutagenesis in Saccharomyces ce
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yal
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
```

GENBANK

ACCESSION NUMBER

- An accession number is a label that used to identify a sequence
  - String of letters and/or numbers that corresponds to a molecular sequence
- An accession number applies to the **complete** database record
  - Remains stable even if updates/ revisions are made to the record

Examples (all for beta globin, *HBB*):

DNA

X02775	GenBank genomic DNA sequence
NG_000007.3	RefSeqGene
rs192792910	dbSNP (single nucleotide polymorphism)

RNA

AA970968.1	An expressed sequence tag (1 of 2,345
NM_000518.4	RefSeq DNA sequence (from a transcript

PROTEIN

NP_000509.1	RefSeq protein
CAA00182.1	GenBank protein
Q14473	SwissProt protein
1YE0|B	Protein Data Bank structure record

http://www.ncbi.nlm.nih.gov/gene/3043

GENBANK

- The version number and GI are unique identifiers for the sequence data within a record
  - Version number
    - A unique number for each sequence submitted
  - Gene Identifier (GI)
    - A unique number for each sequence submitted

http://www.ncbi.nlm.nih.gov/nuccore/1293613/

Saccharomyces cerevisiae TCP1-beta gene, p
(AXL2) and Rev7p (REV7) genes, complete co

GenBank: U49845.1

FASTA Graphics

Go to:

ACCESSION NUMBER

GI

VERSION

```
LOCUS       SCU49845                 5028 bp    DNA     linear
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomy
            Saccharomycetes; Saccharomycetales; Saccharomycet
            Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast require
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Biology, Yale University
            06520, USA
FEATURES             Location/Qualifiers
     source          1..5028
```

GENBANK

- If any change occurs to the sequence data the version number for that sequence is incremented by one decimal and a new GI number is assigned
  - Nucleotide sequence GI number is shown in the VERSION field of the database record
  - Protein sequence GI number is shown in the CDS/db_xref field of a nucleotide database record, and the VERSION field of a protein database record

http://www.ncbi.nlm.nih.gov/nuccore/1293613/

Saccharomyces cerevisiae TCP1-beta gene, p
(AXL2) and Rev7p (REV7) genes, complete co

GenBank: U49845.1

FASTA Graphics

Go to:

ACCESSION NUMBER

GI

VERSION

```
LOCUS       SCU49845                 5028 bp    DNA     linea
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; F[illegible] Ascomycota; Saccharom
            Saccharomyce[illegible]; [illegible]ycetales; Saccharomyce
            Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast require
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Biology, Yale University
            06520, USA
FEATURES             Location/Qualifiers
     source          1..5028
```

GENBANK

GI FOR CDS 1

GI FOR CDS 2

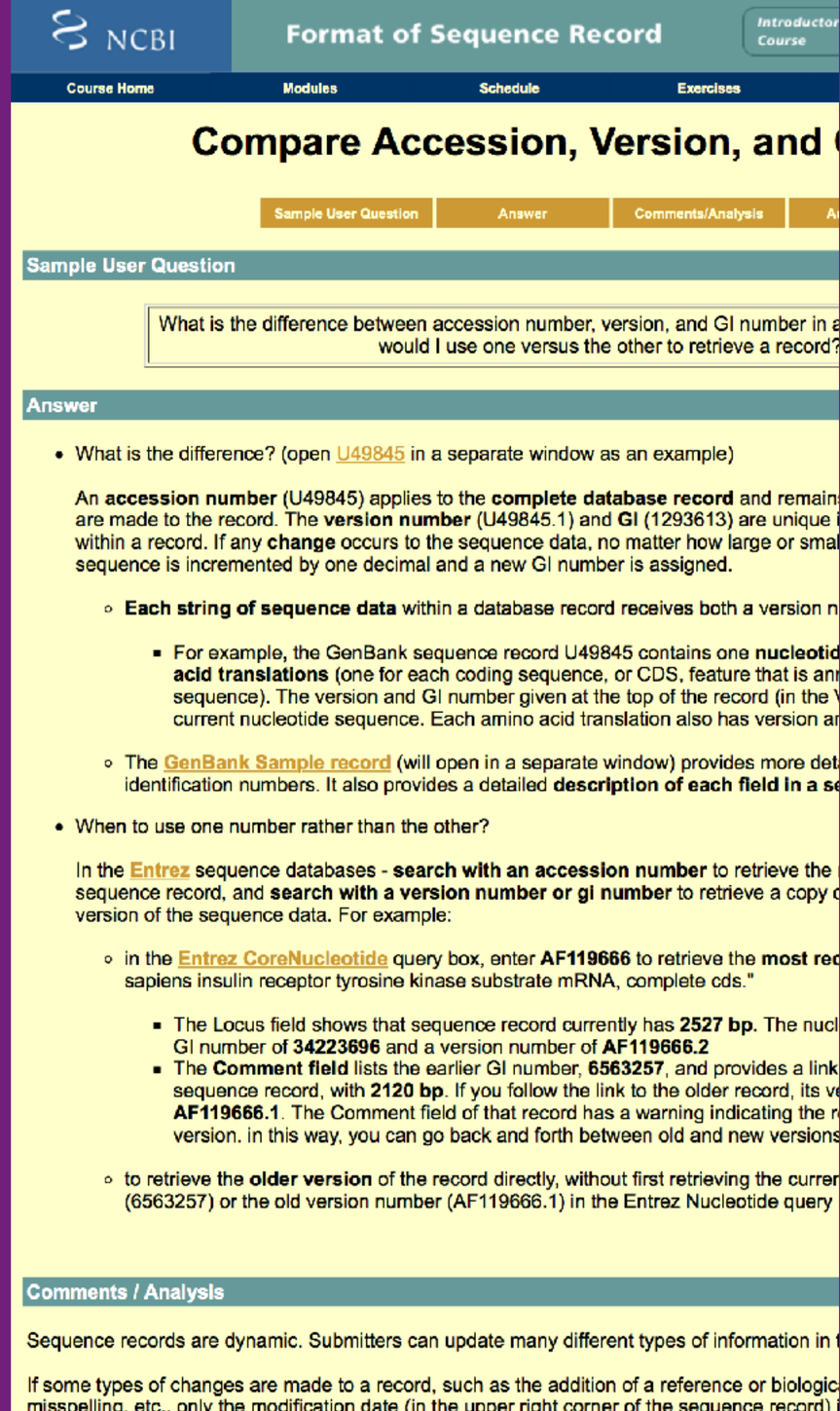
GI FOR CDS 3

- The translated protein sequences have their own GI numbers

```
                 /chromosome="IX"
                 /map="9"
     CDS         <1..206
                 /codon_start=3
                 /product="TCP1-beta"
                 /protein_id="AAA98665.1"
                 /db_xref="GI:1293614"
                 /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                 AEVLLRVDNIIRARPRTANRQHM"
     gene        687..3158
                 /gene="AXL2"
     CDS         687..3158
                 /gene="AXL2"
                 /note="plasma membrane glycoprotein"
                 /codon_start=1
                 /function="required for axial budding pattern of S.
                 cerevisiae"
                 /product="Axl2p"
                 /protein_id="AAA98666.1"
                 /db_xref="GI:1293615"
                 /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                 TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                 VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                 VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                 TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                 YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                 DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                 DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                 NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                 CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                 NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                 SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                 YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                 HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                 VDFSNKSNVNVGQVKDIHGRIPEML"
     gene        complement(3300..4037)
                 /gene="REV7"
     CDS         complement(3300..4037)
                 /gene="REV7"
                 /codon_start=1
                 /product="Rev7p"
                 /protein_id="AAA98667.1"
                 /db_xref="GI:1293616"
                 /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
```

GENBANK

- Why are there two types of sequence identification numbers (GI and VERSION)
  - Different formats; implemented at different times
- When to use one rather than the other?
  - Search with an accession number to retrieve the most recent version of a sequence record
  - Search with a version number or gi number to retrieve a copy of the record that has a specific version of the sequence data



NCBI | Format of Sequence Record | Introductory Course

Course Home | Modules | Schedule | Exercises

Compare Accession, Version, and

Sample User Question | Answer | Comments/Analysis

Sample User Question

What is the difference between accession number, version, and GI number in a
would I use one versus the other to retrieve a record?

Answer

- What is the difference? (open U49845 in a separate window as an example)

  An **accession number** (U49845) applies to the **complete database record** and remain
  are made to the record. The **version number** (U49845.1) and **GI** (1293613) are unique i
  within a record. If any **change** occurs to the sequence data, no matter how large or smal
  sequence is incremented by one decimal and a new GI number is assigned.

  - **Each string of sequence data** within a database record receives both a version n
    - For example, the GenBank sequence record U49845 contains one **nucleotid acid translations** (one for each coding sequence, or CDS, feature that is an sequence). The version and GI number given at the top of the record (in the current nucleotide sequence. Each amino acid translation also has version ar
  - The GenBank Sample record (will open in a separate window) provides more det identification numbers. It also provides a detailed **description of each field in a se**
- When to use one number rather than the other?

  In the Entrez sequence databases - **search with an accession number** to retrieve the sequence record, and **search with a version number or gi number** to retrieve a copy o version of the sequence data. For example:

  - in the Entrez CoreNucleotide query box, enter **AF119666** to retrieve the **most rec** sapiens insulin receptor tyrosine kinase substrate mRNA, complete cds."
    - The Locus field shows that sequence record currently has **2527 bp**. The nucl GI number of **34223696** and a version number of **AF119666.2**
    - The **Comment field** lists the earlier GI number, **6563257**, and provides a link sequence record, with **2120 bp**. If you follow the link to the older record, its ve **AF119666.1**. The Comment field of that record has a warning indicating the r version. in this way, you can go back and forth between old and new versions
  - to retrieve the **older version** of the record directly, without first retrieving the curren (6563257) or the old version number (AF119666.1) in the Entrez Nucleotide query

Comments / Analysis

Sequence records are dynamic. Submitters can update many different types of information in t

If some types of changes are made to a record, such as the addition of a reference or biologic
misspelling, etc., only the modification date (in the upper right corner of the sequence record)

GENBANK

GENBANK DIVISION (PLANT)

ENTRY DATE

IDENTIFIER

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
```

MOLECULE TYPE

SUMMARY

COMPLETE TAXONOMIC INFO FROM NCBI

PUBLICATION

GENBANK

```
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TEQISNDTYKSSVDKTAQITYNCEDLPSWLSEDSSSRTESGEPSSDLLSDANTTLYEN
```

FEATURE RECORDS ARE DIFFERENT DEPENDING ON FEATURE TYPE

GENBANK

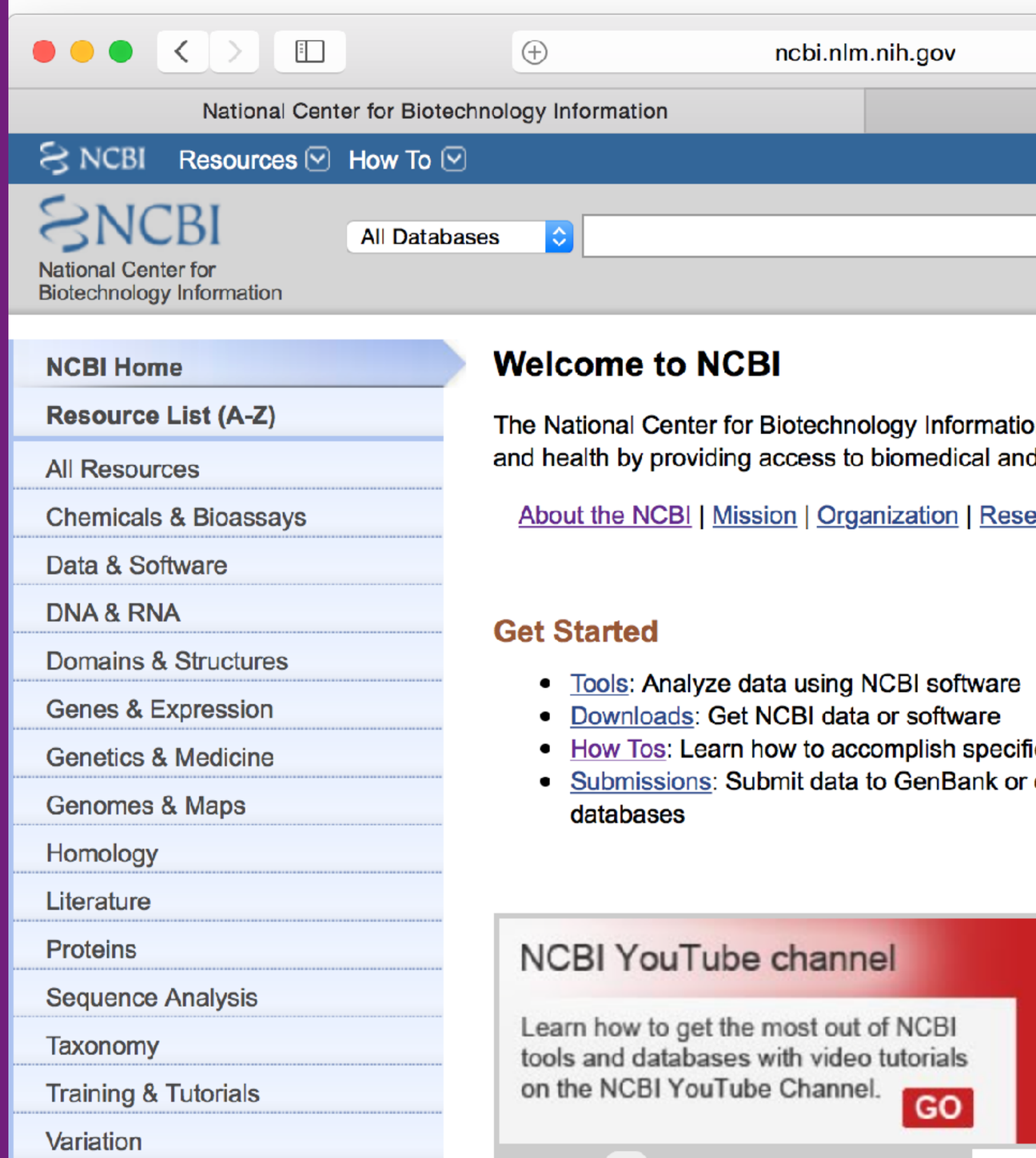
ORIGIN

```
   1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac gga
  61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gta
 121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gca
 181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
 241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
 301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
 361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
 421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
 481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
 541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
 601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
 661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
 721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
 781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
 841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
 901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
 961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
```

THE SEQUENCE DATA

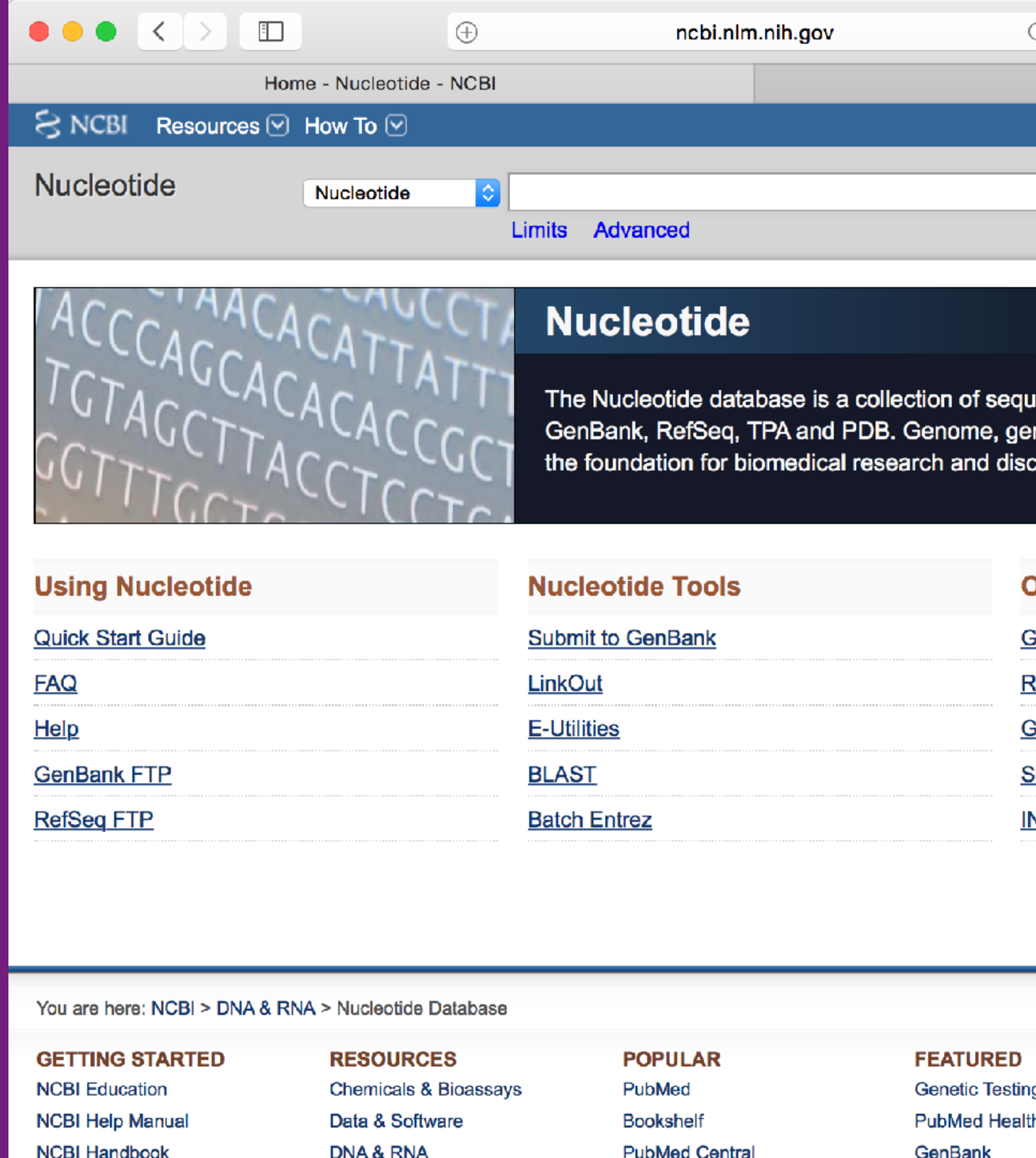
GENBANK

- Access to GenBank
  - Search GenBank for sequence identifiers and annotations with Entrez Nucleotide , which has three divisions:
    - CoreNucleotide (the main collection)
    - dbEST (Expressed Sequence Tags)
    - dbGSS (Genome Survey Sequences)
  - Available as links on records



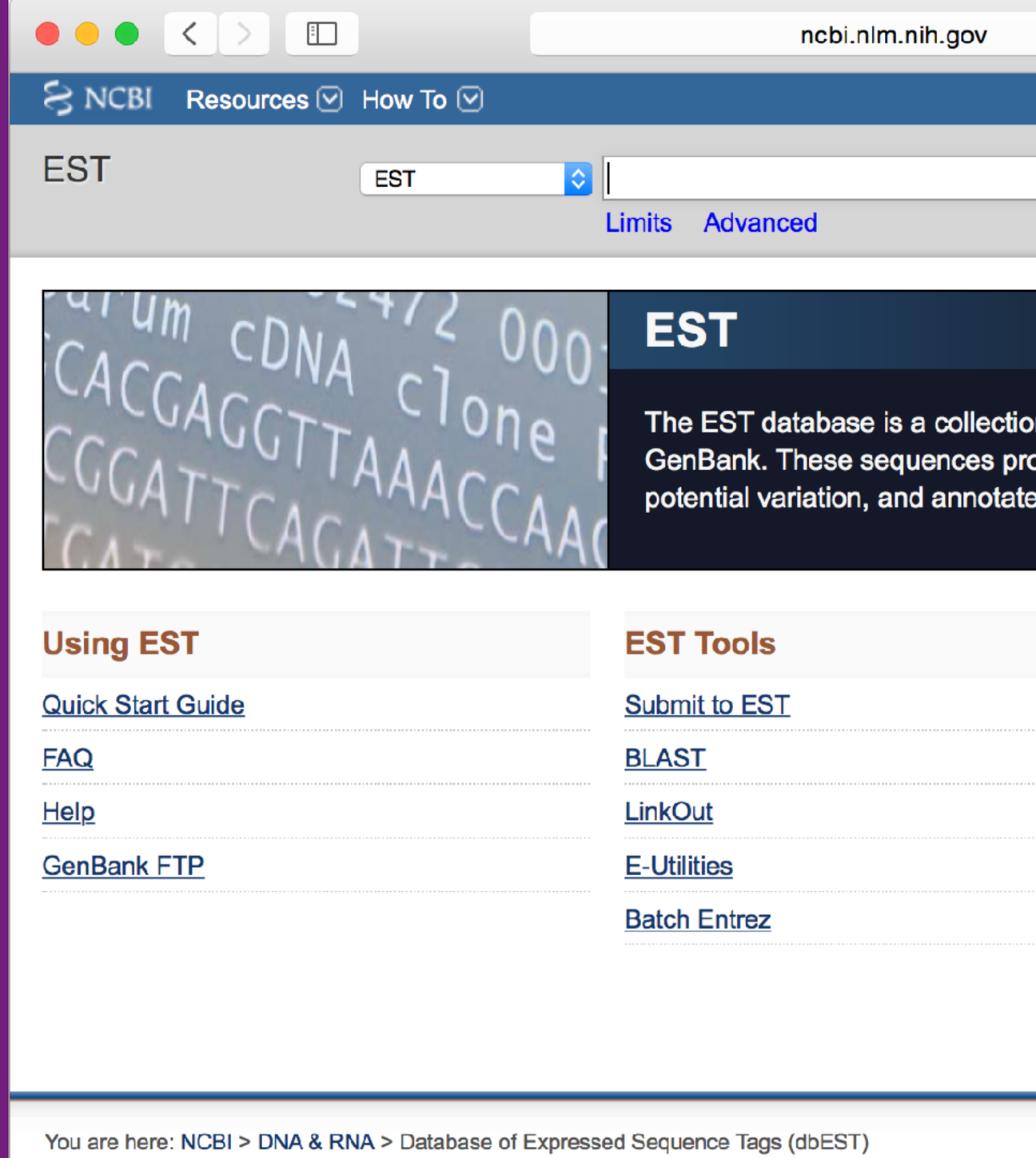
GENBANK

- Nucleotide (Core nucleotide)
  - A collection of sequences from several sources
    - GenBank, RefSeq, TPA and PDB
    - Genome, gene and transcript sequence data



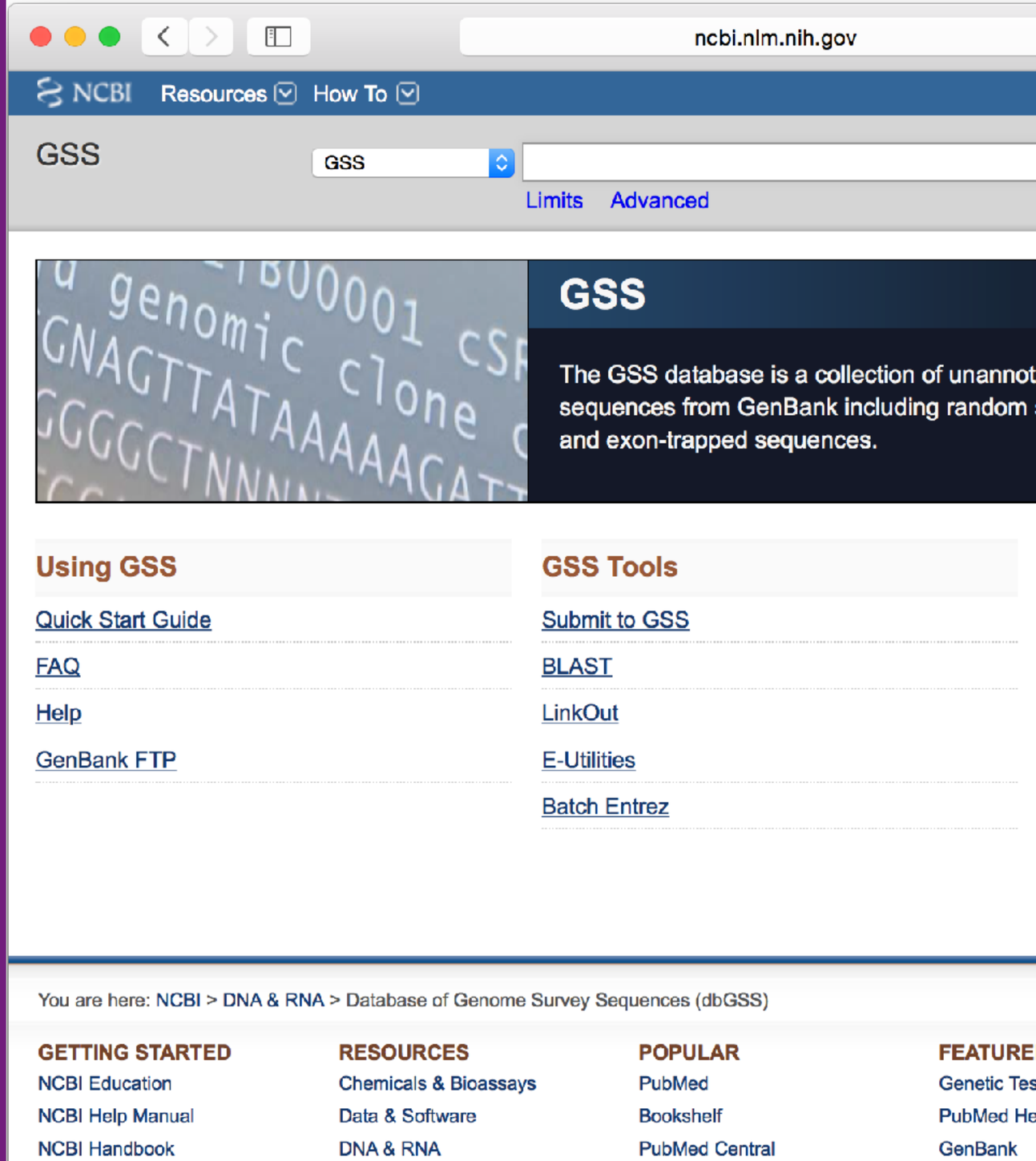
GENBANK

- dbEST (Expressed Sequence Tags)
  - Short single-read transcript sequences from GenBank
- ESTs used to
  - Evaluate gene expression
  - Find potential variation
  - Annotate genes



GENBANK

- Genomic Sequence Survey (GSS)
  - The GSS is similar to the EST division
  - Exception that most of the sequences are genomic in origin

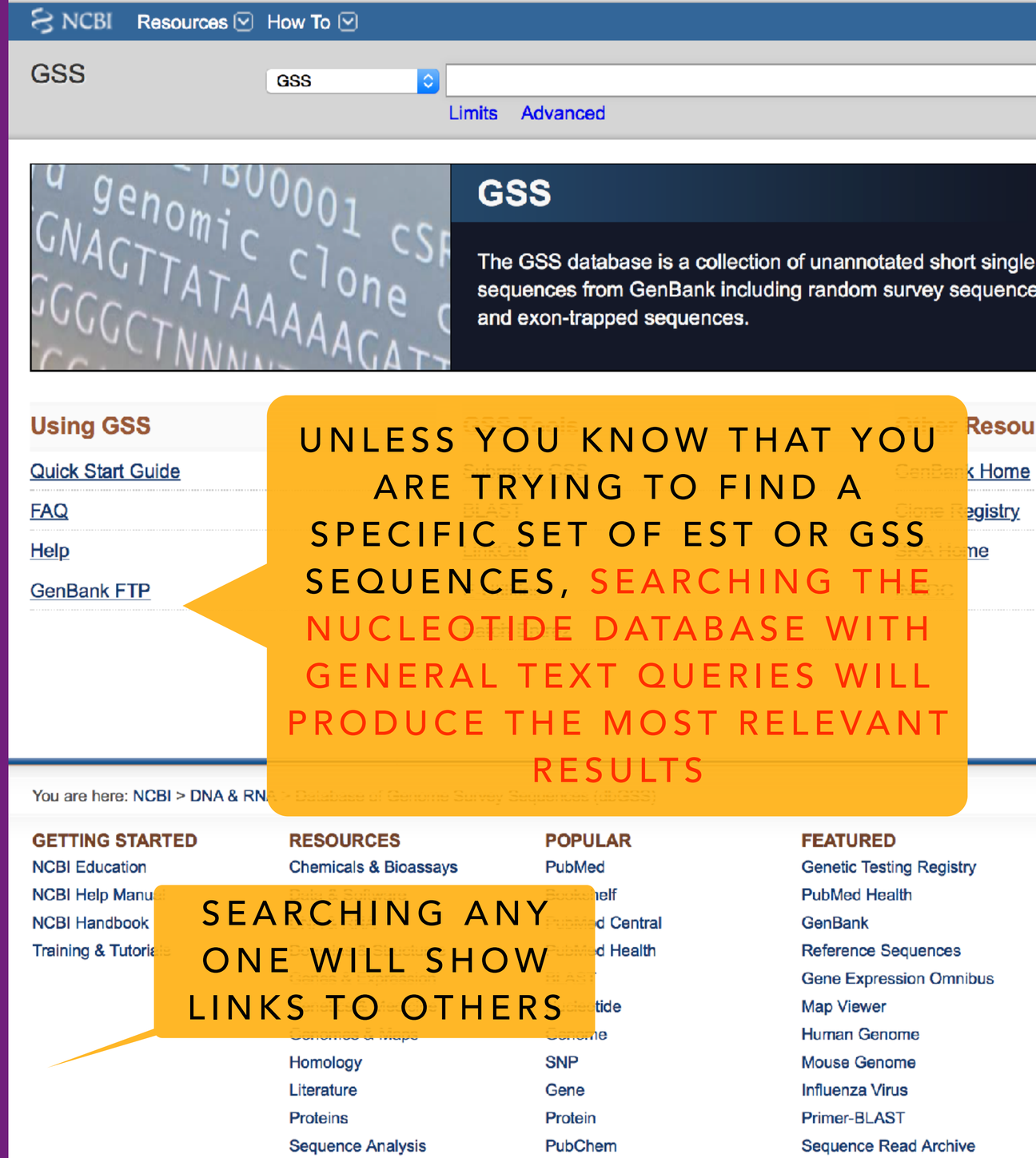


GENBANK

WHICH SHOULD I USE?

- The Nucleotide, Genome Survey Sequence (GSS), and Expressed Sequence Tag (EST) database all contain nucleic acid sequences

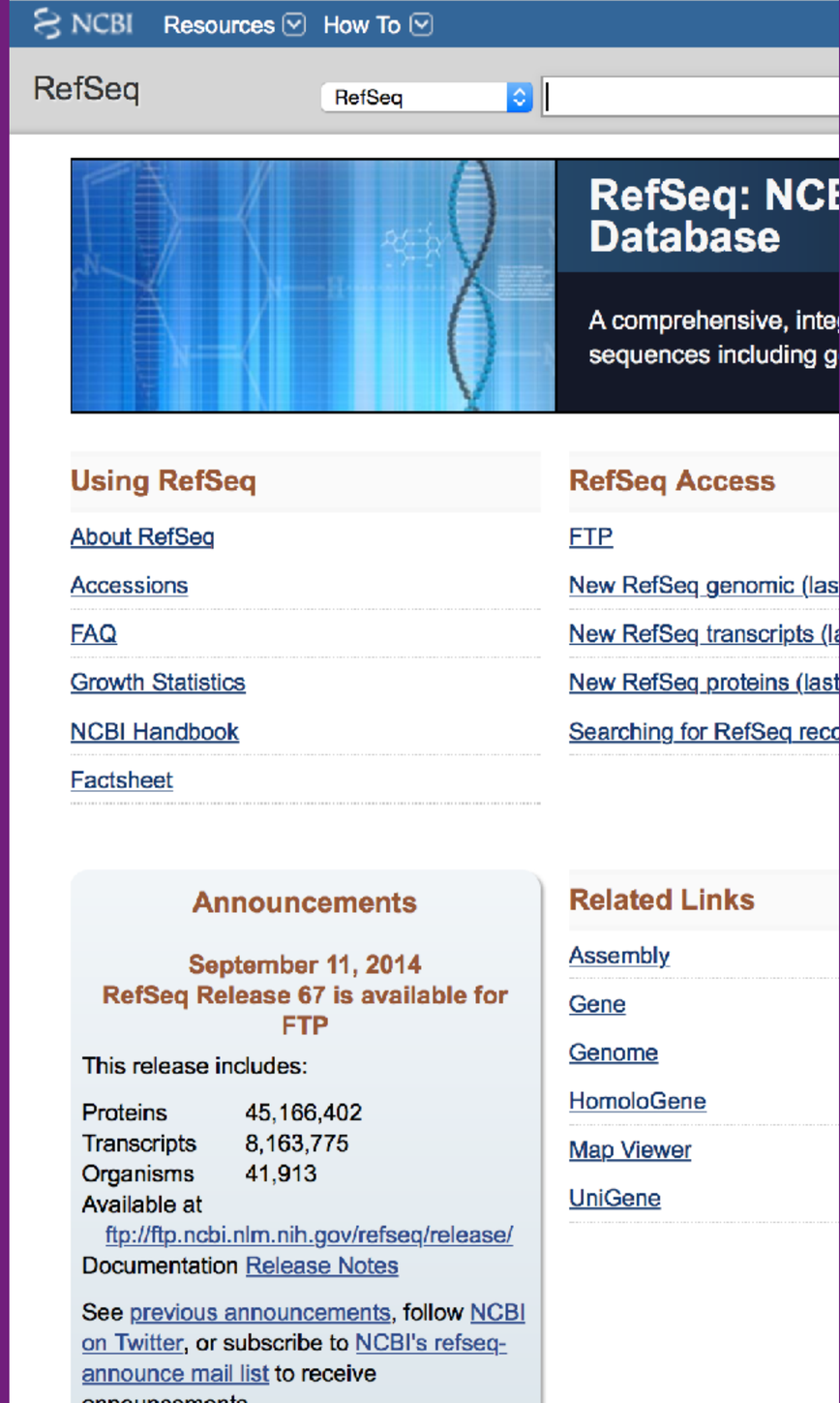
- Data in GSS and EST are from two large bulk sequence divisions of GenBank GSS and EST data are typically uncharacterized, short genomic (GSS) or cDNA (EST) sequences



RefSeq

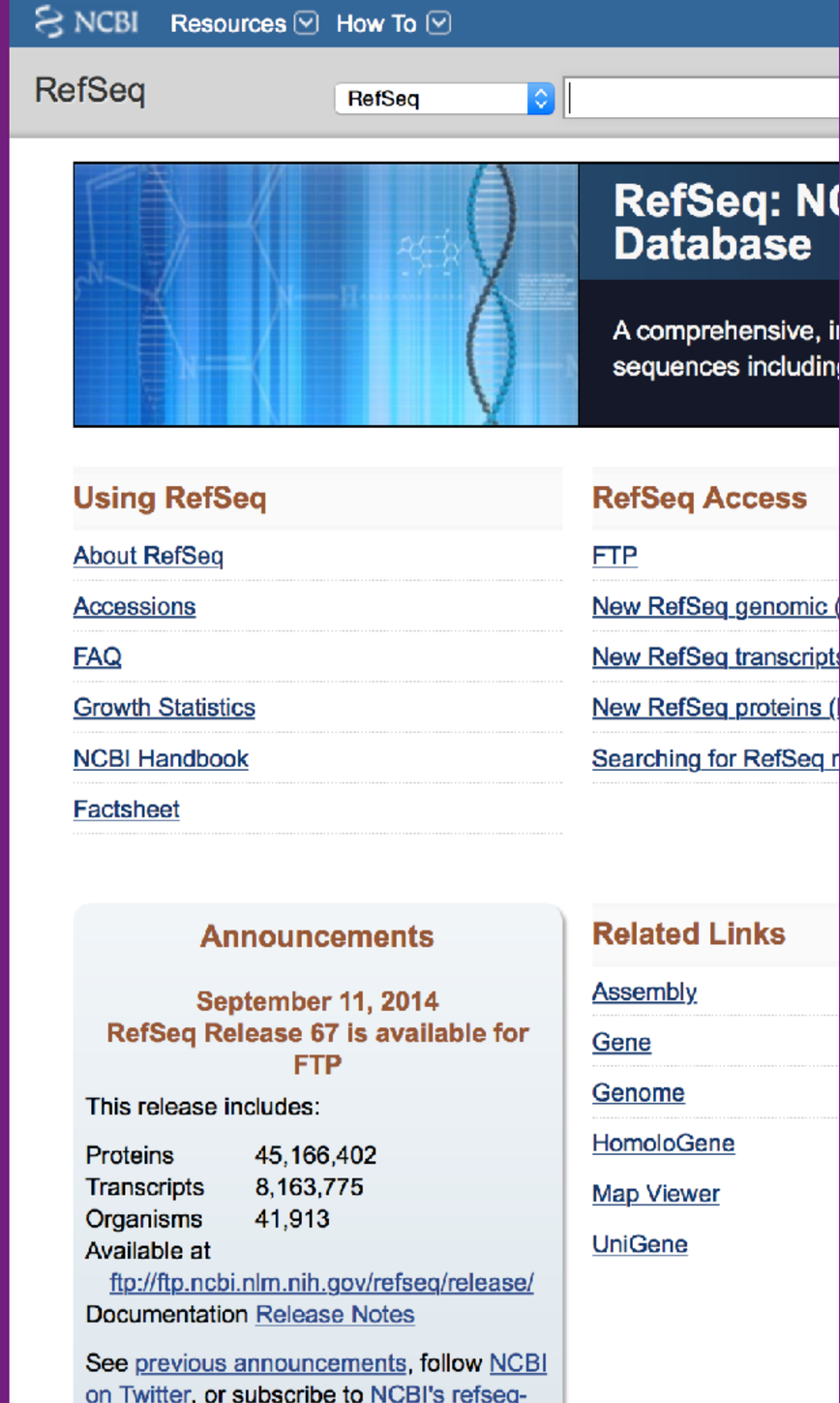
REFSEQ

- RefSeq sequences form a foundation for medical, functional, and diversity studies
- They provide a stable reference
  - genome annotation
  - gene identification and characterization
  - mutation and polymorphism analysis (especially RefSeqGene records)
  - expression studies
  - comparative analyses



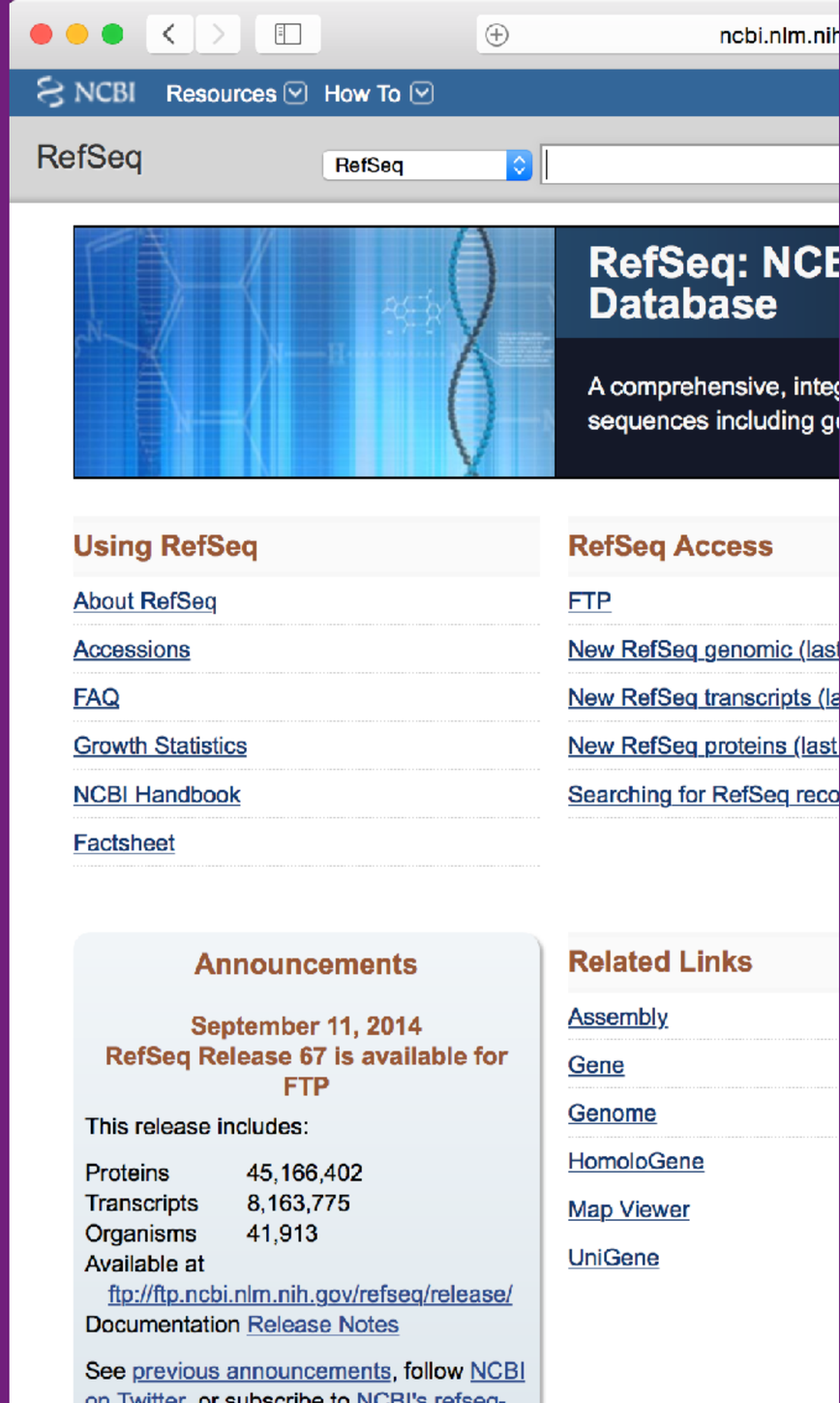
REFSEQ

- The Reference Sequence (RefSeq) collection
  - comprehensive and integrated
  - non-redundant
  - well-annotated
- Composed of
  - Genomic DNA, RNA Transcripts, Proteins



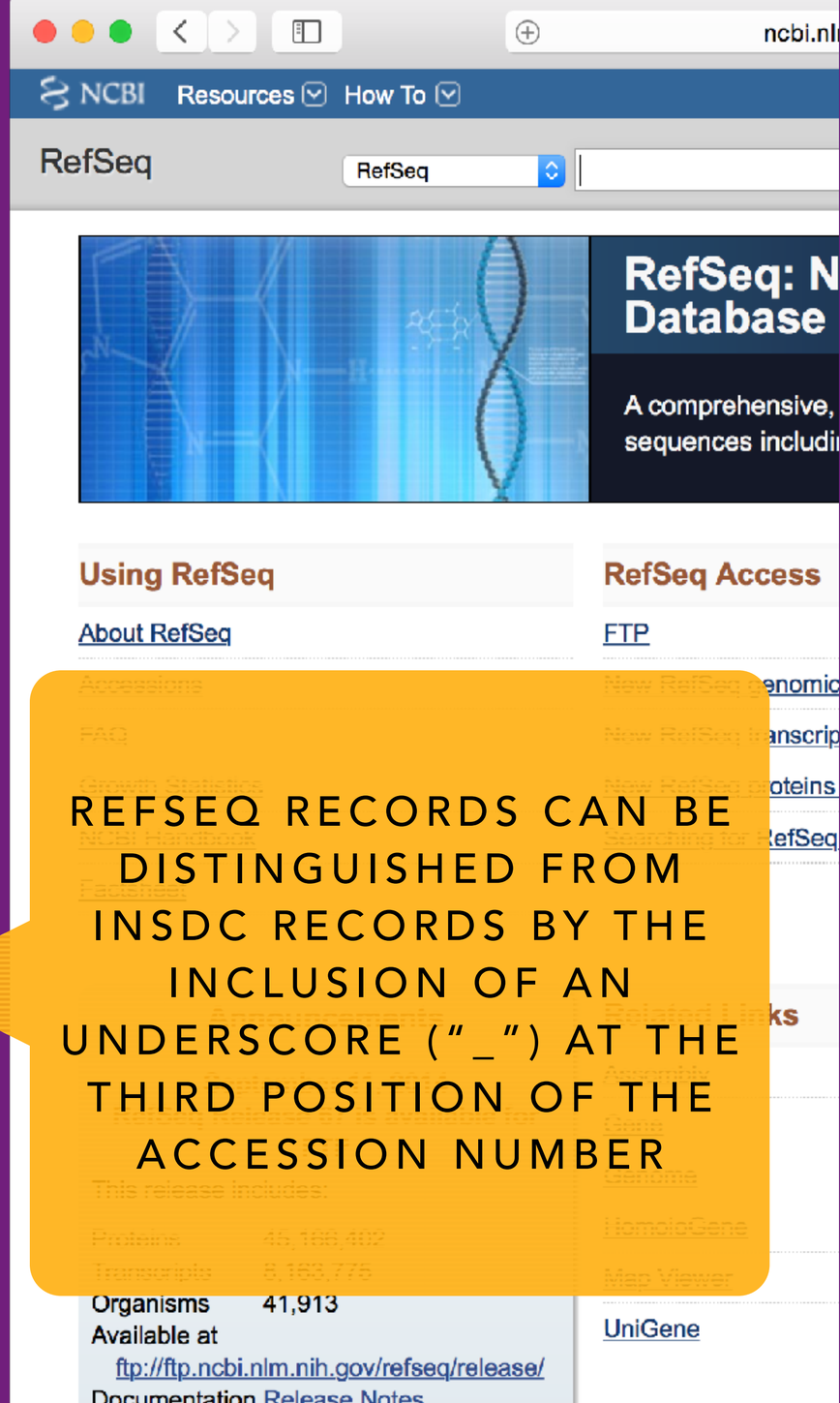
REFSEQ

- Each RefSeq is constructed wholly from sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC)
- RefSeq is a synthesis of information integrated across multiple sources at a given time
  - Ongoing curation by NCBI staff
  - Updated to reflect current knowledge



REFSEQ

- 12,000 distinct taxonomic identifiers
  - Includes sequences from plasmids, organelles, viruses, archaea, bacteria, and eukaryotes
- Every sequence has a
  - Stable accession number
  - Version number
  - Integer identifier (gi)



REFSEQ RECORDS CAN BE DISTINGUISHED FROM INSDC RECORDS BY THE INCLUSION OF AN UNDERSCORE ("_") AT THE THIRD POSITION OF THE ACCESSION NUMBER

REFSEQ

NZ_ [b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_ [c]	mRNA	Predicted model
XR_ [c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly

- The RefSeq accession prefix has an implied meaning in terms of the type of molecule it represents

REFSEQ

Accession Format	Molecule Type
NC_123456	Complete Genome Complete Chromosome Complete Sequence
NG_123456	Genomic Region
NM_123456	mRNA
NR_123456	non-coding RNA
NP_123456	Protein
NT_123456	Genomic Contig (from BACs)
NW_123456	Genomic Contig (from WGS)
XM_123456	mRNA (taken from genomic seq)
XR_123456	RNA (taken from genomic seq)
XP_123456	Protein (taken from genomic seq)

REFSEQ

STATUS

Code	Description
MODEL	The RefSeq record is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs.
INFERRED	The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
PREDICTED	The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.
PROVISIONAL	The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff.
REVIEWED	The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information

- RefSeq records appear similar in format to GenBank records
- COMMENT field that indicates the RefSeq status and the INSDC source of the sequence information

REFSEQ

GCNT2 glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group) [*Homo sapiens* (human)]

Gene ID: 2651, updated on 3-Oct-2014

Summary

Official Symbol	GCNT2 provided by HGNC
Official Full Name	glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group) provided by HGNC
Primary source	HGNC:HGNC:4204
Locus tag	RP11-421M1.1
See related	Ensembl:ENSG00000111846; HPRD:02695; MIM:600429; Vega:OTTHUMG00000014244
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	IL; GCAT; IGNT; ULG3; GCNT5; GCNT2C; NACGT1; NACGT1; CTRCT13;

FEATURES OF A REFSEQ RECORD

Table of contents
Summary
Genomic context
Genomic regions, transcripts, and products
Expression
Bibliography
Phenotypes
Variation
Pathways from BioSystems
Interactions
General gene information
Markers, Related pseudogene(s), Clone Names, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences

REFSEQ

```
     PUBMED     458154
COMMENT         REVIEWED REFSEQ: This record has been curated by NCBI staff. The
                reference sequence was derived from DB208581.1, AK090483.1 and
                AL832719.1.
                On Jun 8, 2008 this sequence version replaced gi:85790416.

                Summary: This gene encodes the enzyme responsible for formation of
                the blood group I antigen. The i and I antigens are distinguished
                by linear and branched poly-N-acetyllactosaminoglycans,
                respectively. The encoded protein is the I-branching enzyme, a
                beta-1,6-N-acetylglucosaminyltransferase responsible for the
                conversion of fetal i antigen to adult I antigen in erythrocytes
                during embryonic development. Mutations in this gene have been
                associated with adult i blood group phenotype. Alternatively
                spliced transcript variants encoding different isoforms have been
                described. [provided by RefSeq, Jul 2008].

                Transcript Variant: This variant (1) contains a different 5' end
                exon compared to variant 2. The encoded isoform (A) is longer and
                has a distinct N-terminus, compared to isoform B.

                Publication Note:  This RefSeq record includes a subset of the
                publications that are available for this gene. Please see the Gene
                record to access additional publications.

                ##Evidence-Data-START##
                Transcript exon combination :: AK291767.1 [ECO:0000332]
                RNAseq introns              :: single sample supports all introns
                                               ERS025082, ERS025087 [ECO:0000348]
                ##Evidence-Data-END##
                COMPLETENESS: complete on the 3' end.
PRIMARY         REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
                1-465               DB208581.1         2-466
                466-4520            AK090483.1         471-4525
                4521-4540           AL832719.1         4201-4220
FEATURES             Location/Qualifiers
```

siRNA and shRNA [ExactAnti
Order full-length cDNA clon [Gene
TaqMan® Gene Expression [Life
TaqMan® Genotyping Assa [Life
PCR/Sanger Sequencing P Assays [Life
Ambion siRNA [Life
TaqMan® Copy Number Va Assays [Life
imaGenes
UCSC Genome Browser [UCSC Genc

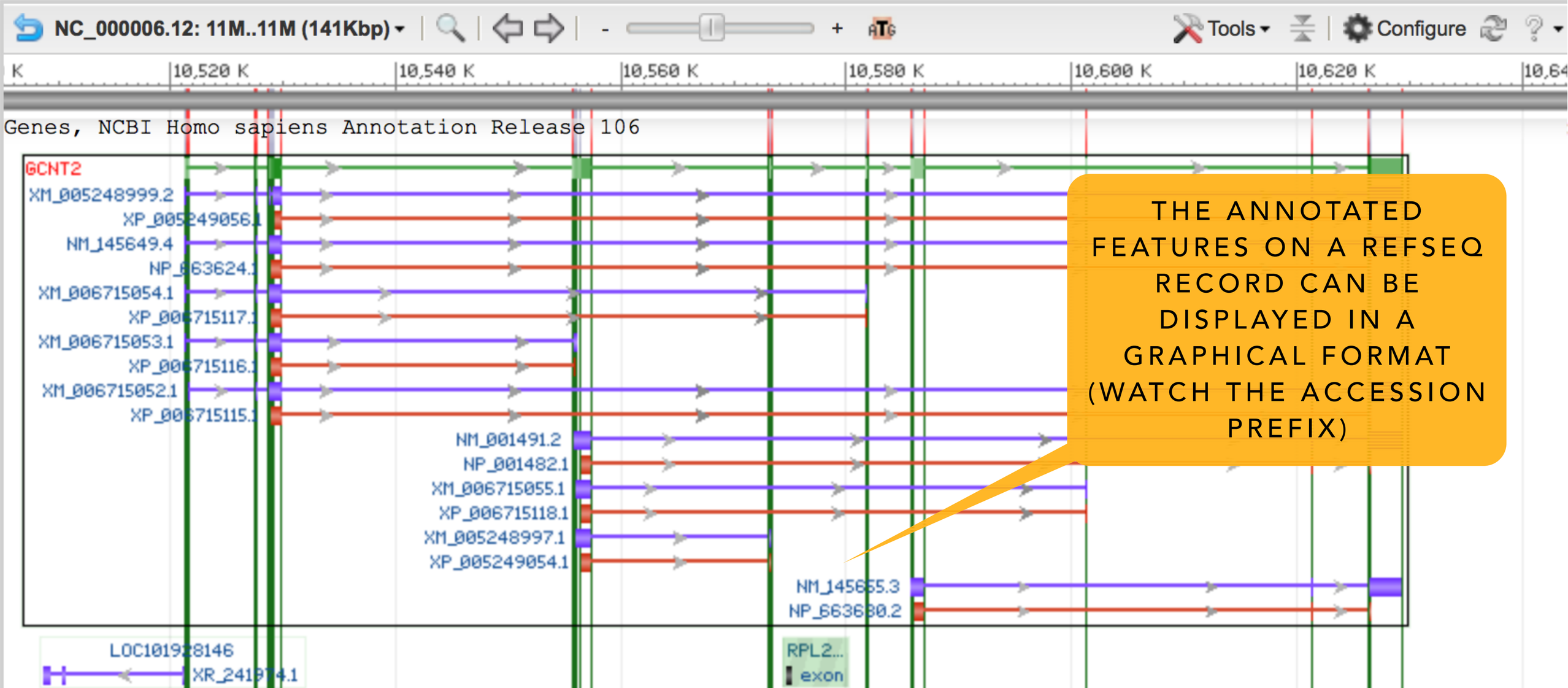
Related information

Related Sequences
BioSystems
CCDS
Components (Core)
Components (EST)
Full text in PMC

THE COMMENT AND PRIMARY SECTIONS

THE GENE SUMMARY IS PROVIDED FOR REFSEQS WITH A REVIEWED STATUS ONLY

REFSEQ



THE ANNOTATED FEATURES ON A REFSEQ RECORD CAN BE DISPLAYED IN A GRAPHICAL FORMAT (WATCH THE ACCESSION PREFIX)

REFSEQ

- RefSeq records can be accessed
  - By direct query
  - BLAST, FTP download, or indirectly through links provided from several NCBI resources, including Gene, Genome, BioProject, and Map Viewer

REFSEQ

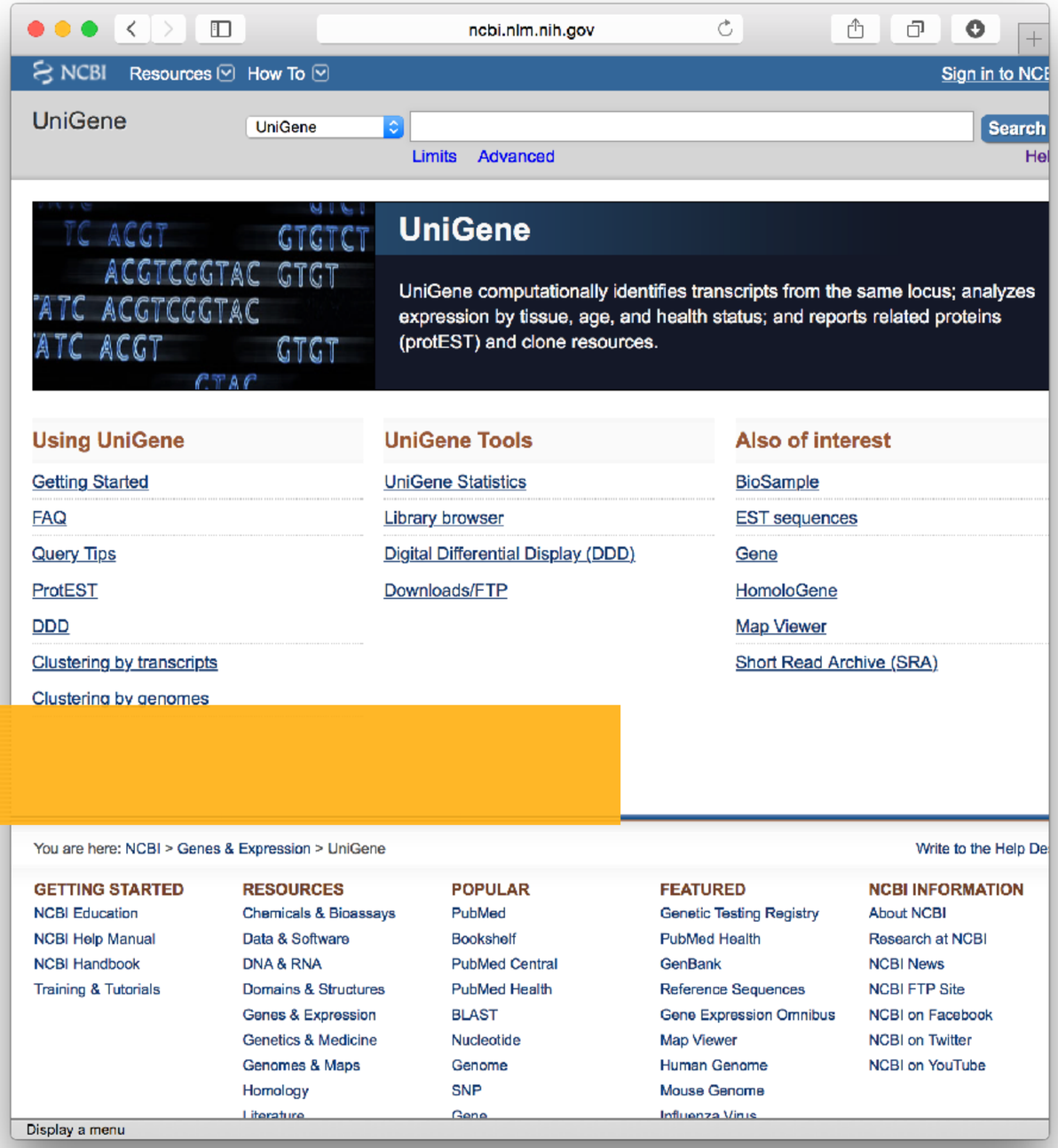
GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

- What should I use? RefSeq vs GenBank

Notable Databases

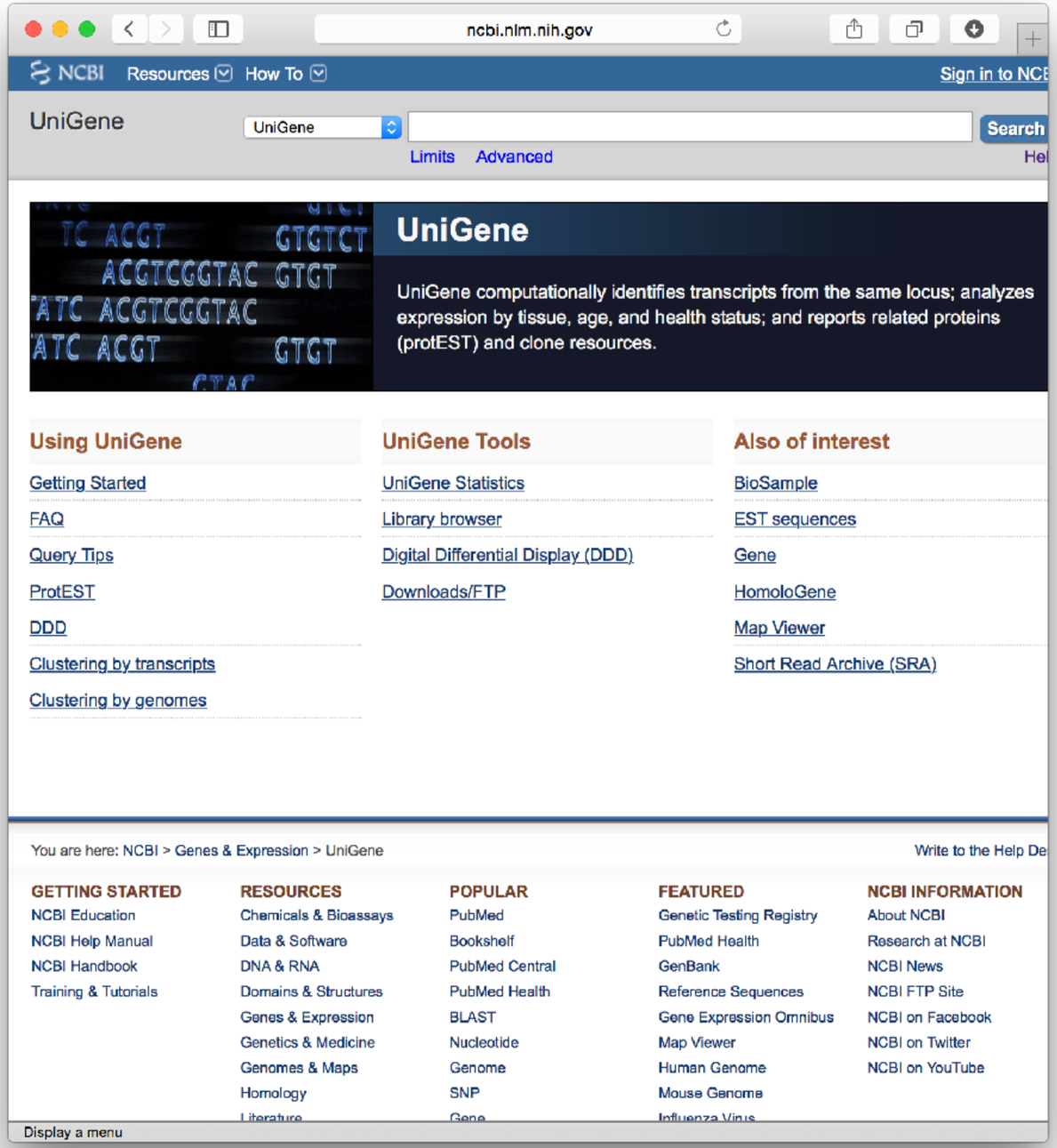
UNIGENE

- UniGene
  - Put together mRNA from combination of ESTs
  - Computationally determined
  - Sequence alignment available
  - Unstable - It is regularly rebuilt



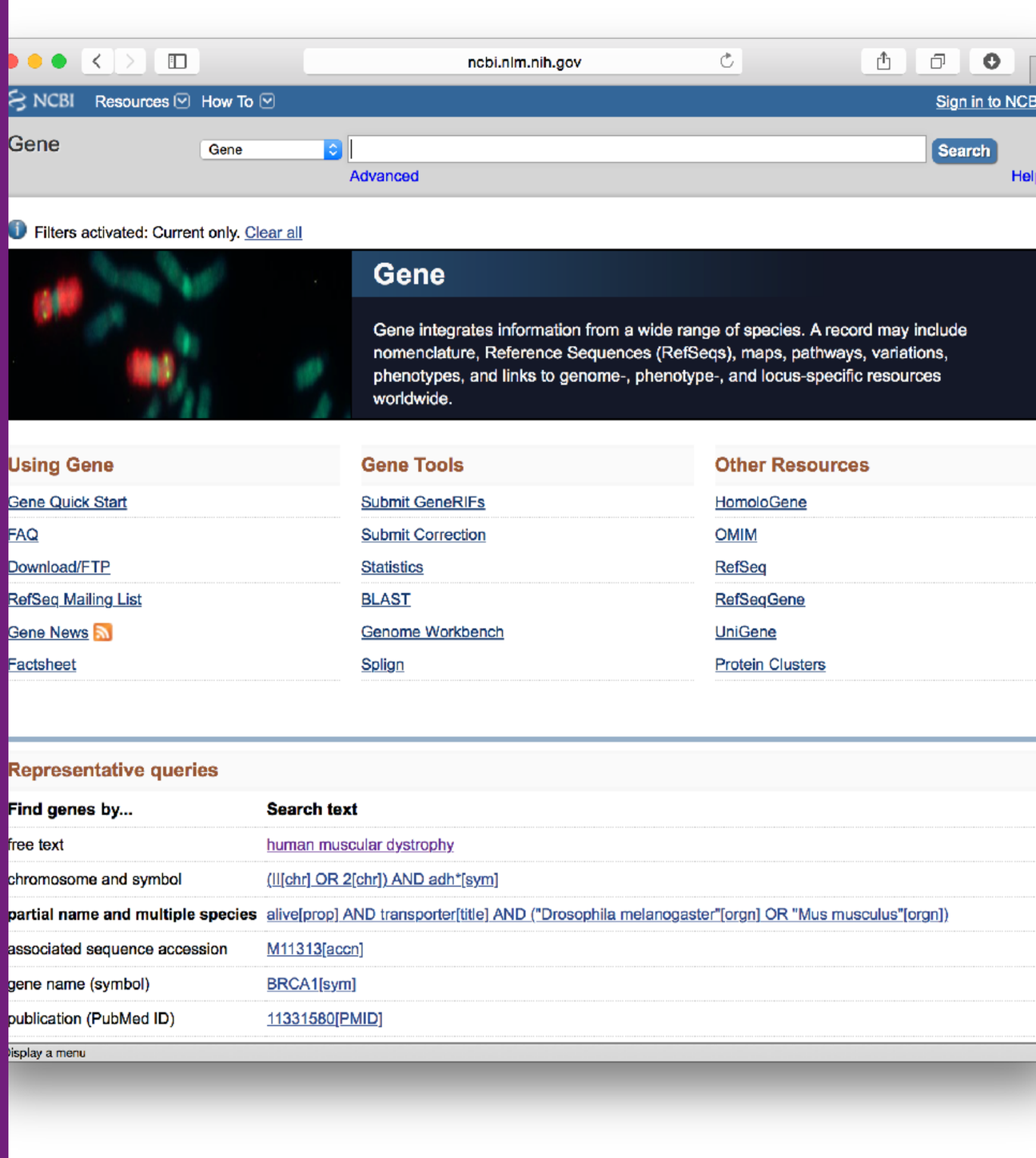
UNIGENE

- UniGene
  - Analyzes expression by tissue, age, and health status
  - Reports related proteins (protEST) and clone resources



GENE

- Gene
  - Integrated Access to Genes of Genomes in the Reference Sequence Collection



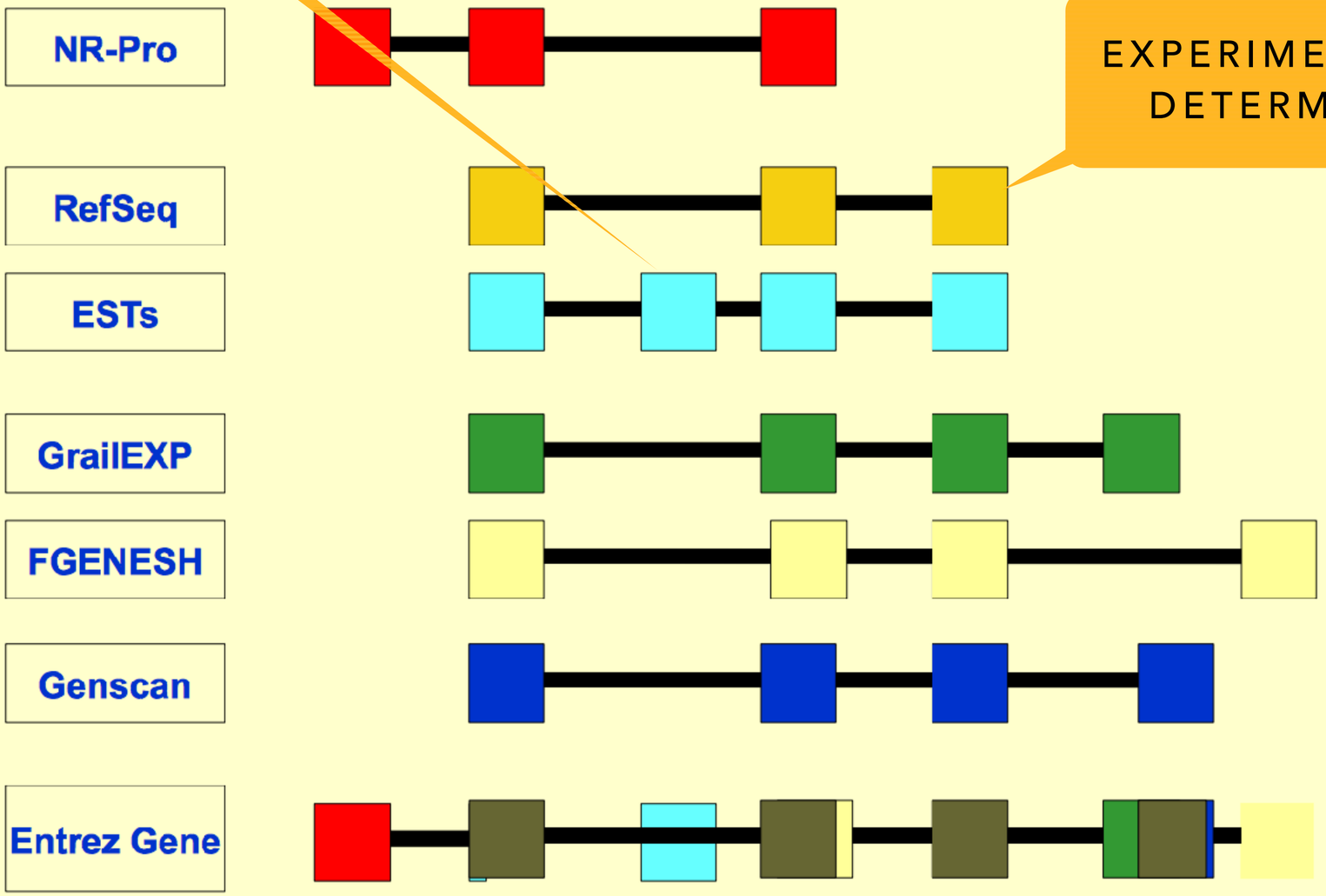
GENE

EXON FROM A DIFFERENT TISSUE

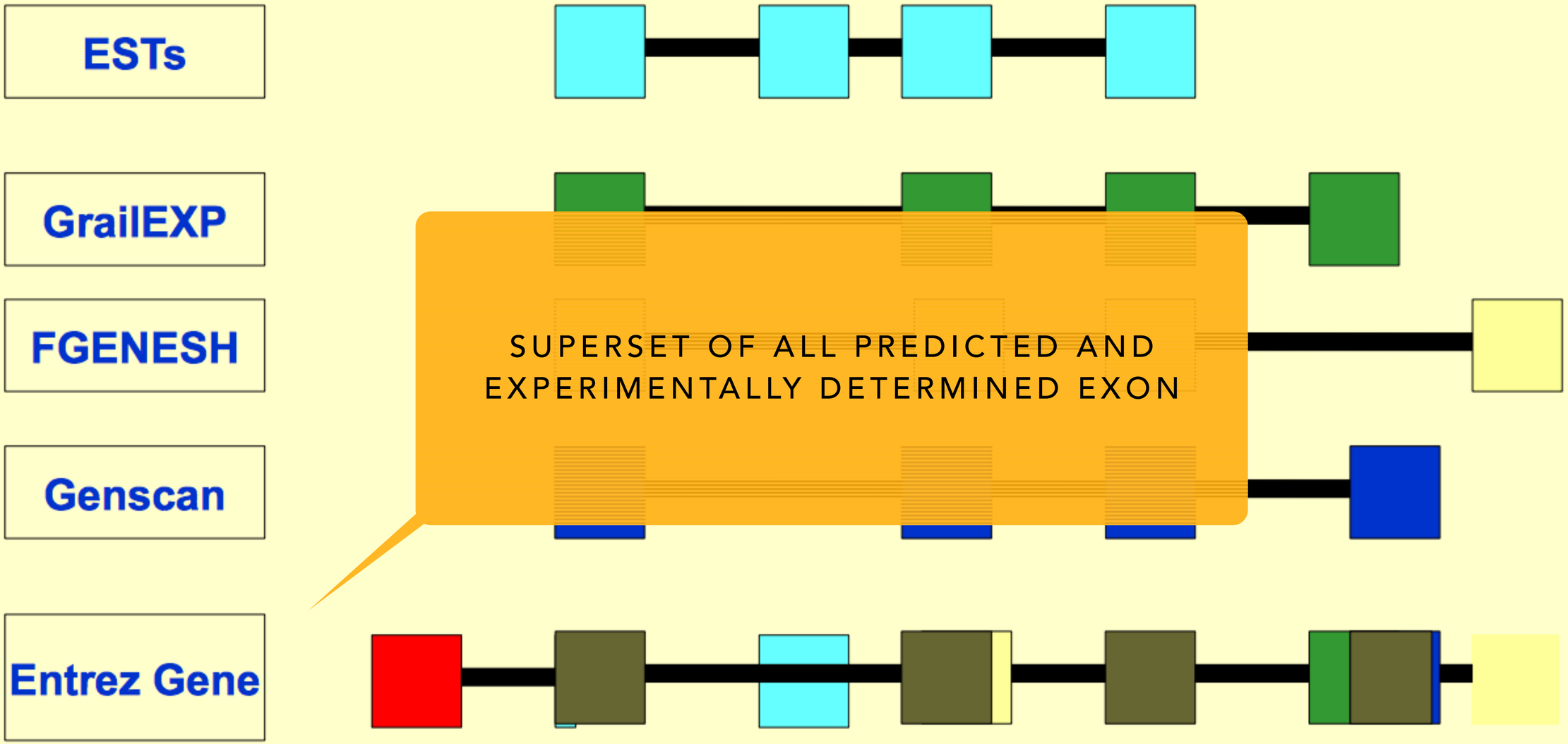
PROTEIN SEQUENCE FROM THE GENE FULL LENGTH CDNA DOESN'T GO TO 5' END

EXPERIMENTALLY DETERMINED

PREDICT EXONS

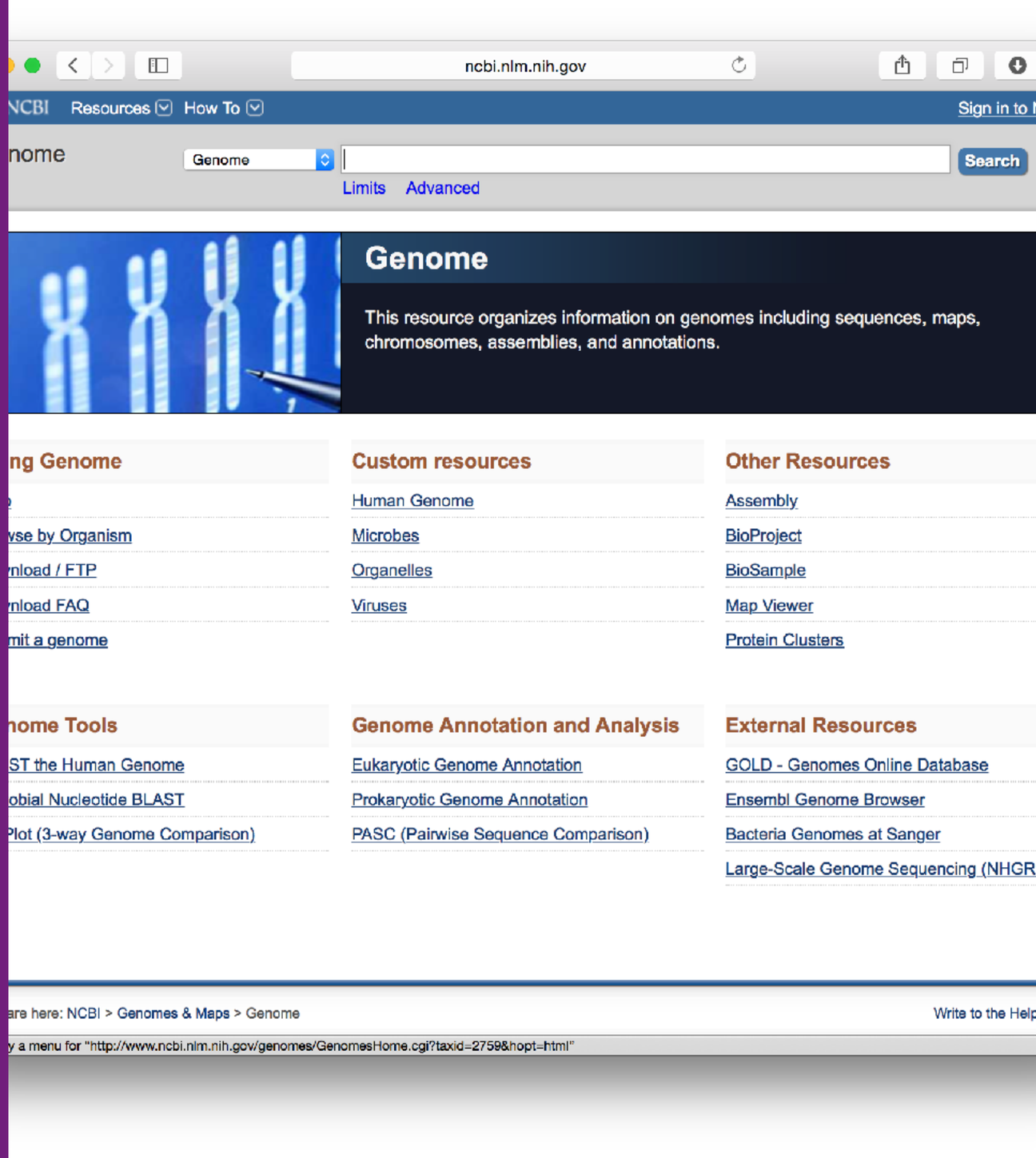


GENE


ESTs
GrailEXP
FGENESH
Genscan
Entrez Gene
SUPERSET OF ALL PREDICTED AND EXPERIMENTALLY DETERMINED EXON

NCBI

- Genome
  - Contains sequence and map data from the whole genomes of 1000+ species or strains
  - Represent completely sequenced genomes
  - Store in-progress sequencing



NCBI

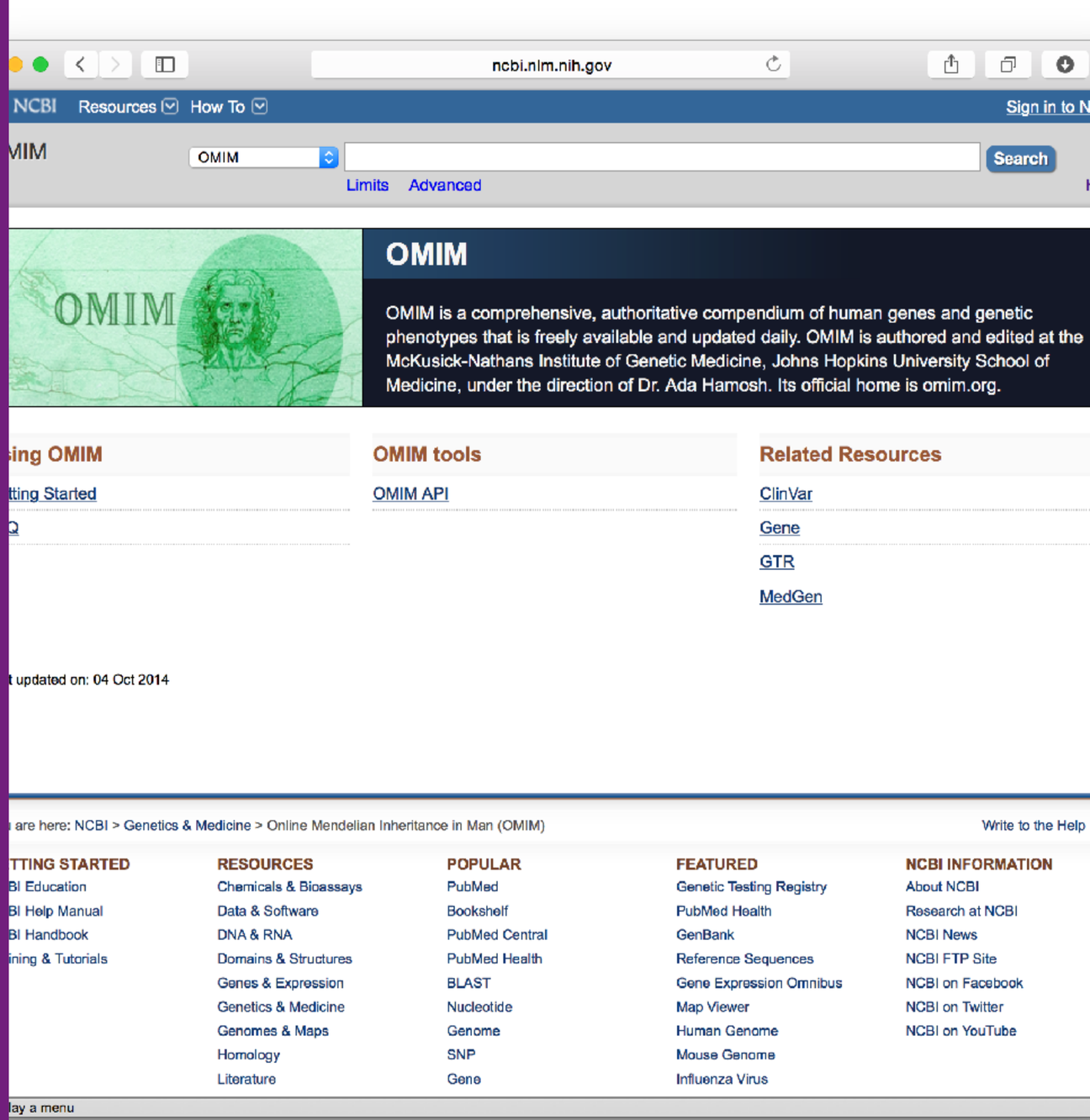
Items 1 - 100 of 1680 << First < Prev Page 1 of 17 Next > Last >>

Organism/Name	BioSample	BioProject	Group -- All	SubGroup -- All Euk	Assembly	Level	Size (Mb)	GC%	Chrs
Acanthamoeba castellanii	SAMN00000060	PRJNA20303	Protists	Other Protists	GCA_000193105.1		46.7146	58.40	-
Acanthamoeba castellanii str. Neff	SAMN02953809	PRJNA193615	Protists	Other Protists	GCA_000313135.1		42.0198	58.40	-
Acanthisitta chloris	SAMN02318032	PRJNA253841	Animals	Birds	GCA_000695815.1		1035.88	41.60	-
Acanthoscurria geniculata	SAMN02720822	PRJNA222716	Animals	Other Animals	GCA_000661875.1		7178.4	0.00	-
Aciculosporium take MAFF-241224	SAMN02981340	PRJNA67241	Fungi	Ascomycetes	GCA_000222935.2		58.8364	40.10	-
Acromyrmex echinatior	SAMN02953789	PRJNA62733	Animals	Insects	GCA_000204515.1		295.945	33.70	-

- Search by organism for entire genome
  - Not individual sequences

NCBI

- Online Mendelian Inheritance in Man (OMIM)
  - An Online Catalog of Human Genes and Genetic Disorders
  - Comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily
- Links to everything known about a given disease across NCBI databases

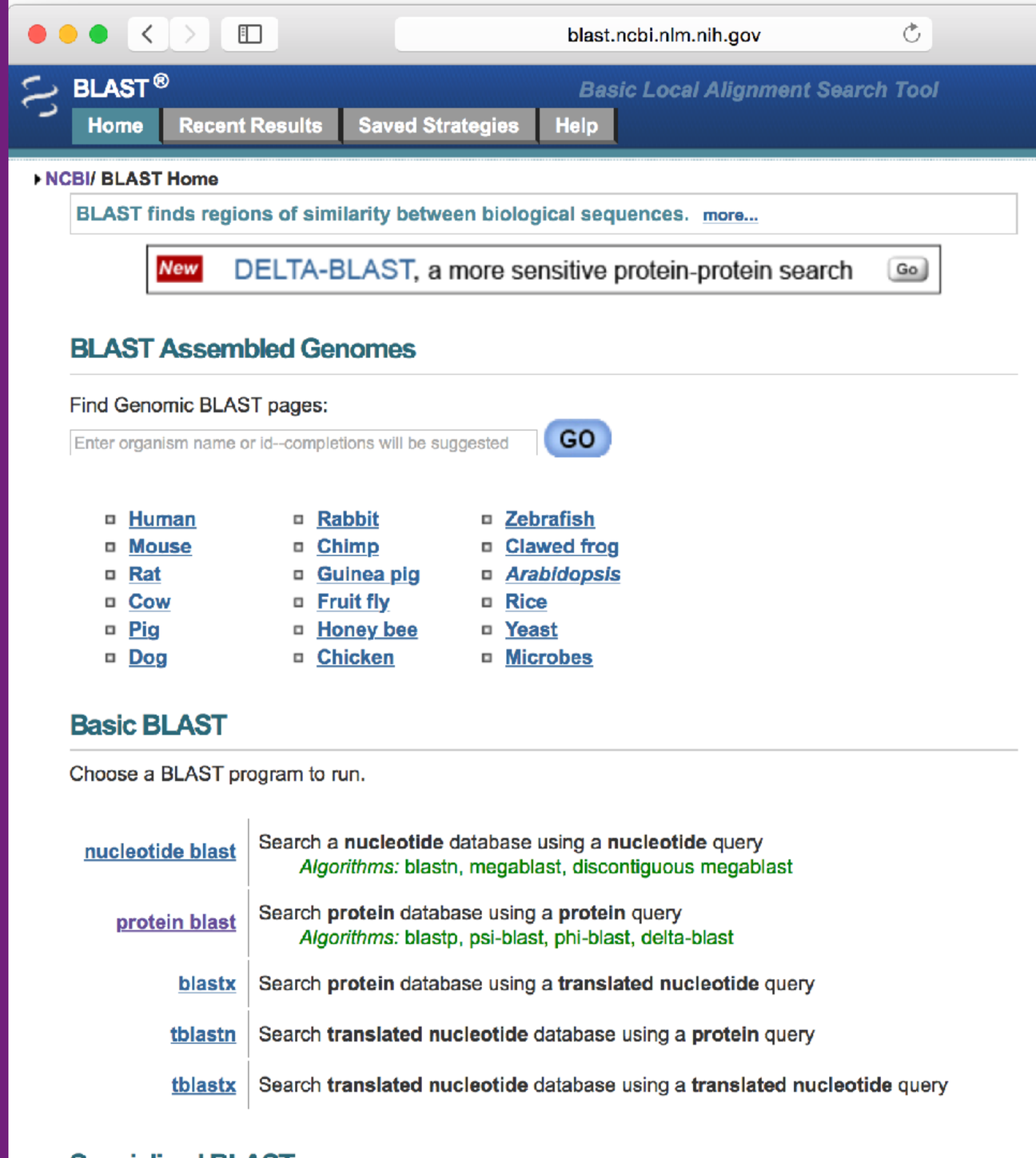


LAB: NAVIGATING NCBI

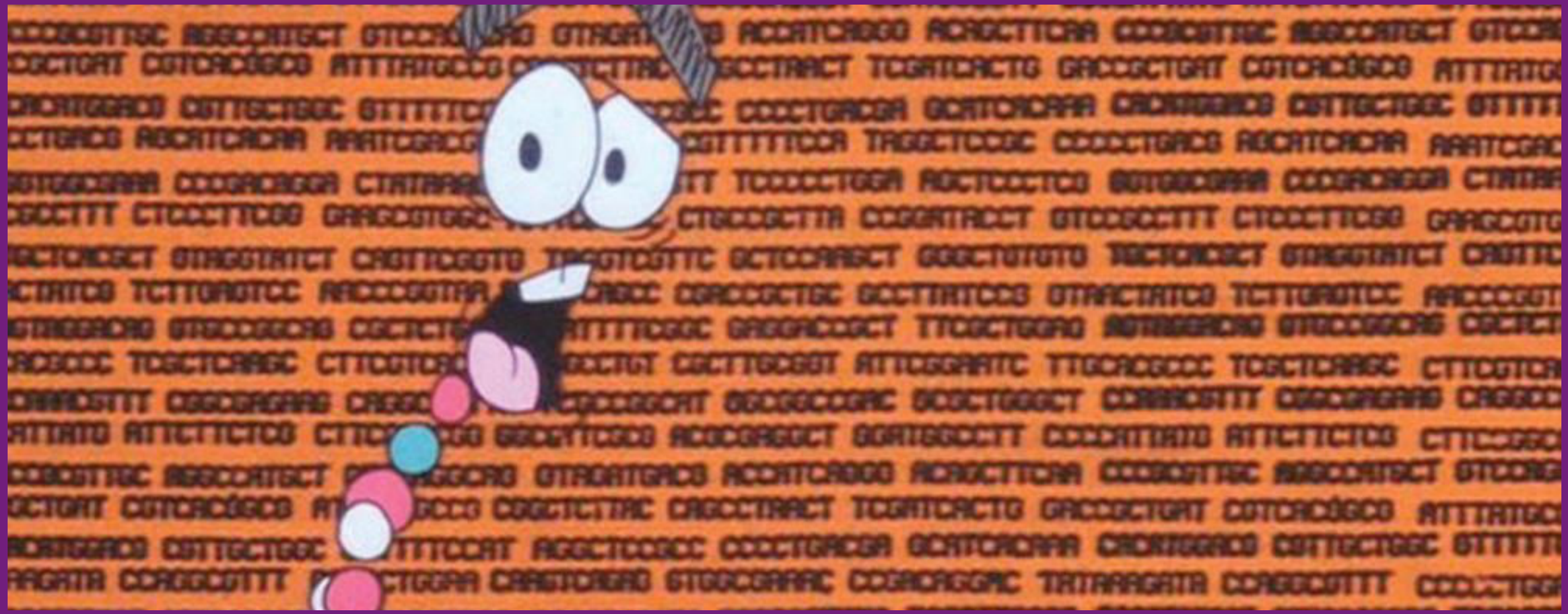
Jurassic Park

JURASSIC PARK

- BLAST
  - BLAST finds regions of similarity between biological sequences
  - Many options and variants (which we will discuss next week)



JURASSIC PARK



- Mr. DNA in Jurassic Park tour video claims they filled in the gaps with amphibian DNA

JURASSIC PARK

- Mark S. Boguski (NCBI at the time) decided to find out if it was true

Book Report

A Molecular Biologist Visits *Jurassic Park*

Mark S. Boguski
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

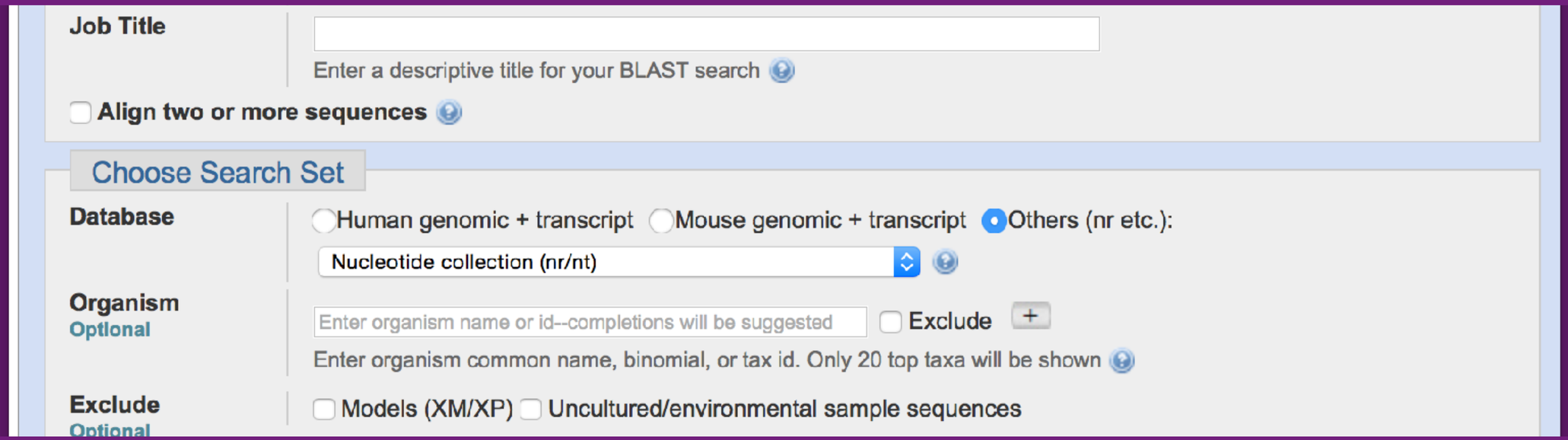
What do you get when you mix dinosaurs and DNA, chaos and computers, Frankenstein and fractals, vortex briefly and then divide into several dozen movie scene-sized aliquots? The answer is Michael Crichton's techno-thriller *Jurassic Park* (1990, Knopf, New York) that has recently been released in paperback. Crichton is a Harvard Medical School graduate who achieved popular success as a writer with his book *The Andromeda Strain* and went on to direct hit movies such as *Westworld* and *Coma*. Having greatly enjoyed his earlier work, I eagerly awaited an opportunity to read *Jurassic Park* which is based on the premise that molecular biotechnology will permit the resurrection of extinct animals from fossilized DNA.

capital), a fictional biotechnology firm called InGen create
biological Disneyland on an island off the coast of Costa Ri
Merchandizing possibilities rival or exceed those of t
Teenage Mutant Ninja Turtles.

Early in the book, InGen's chief scientist, Dr. Henry Wu,
giving a tour of the laboratory facilities which contain thr
Cray® XMP supercomputers and automated gene sequence
which are called "hoods" (an amusing allusion to Lee Hood
Applied Biosystems that will be lost on the general reader).
one stop on the tour (p. 103), we are shown an actual samp
of dinosaur DNA sequence

```
1  gcgttgctgg cgtttttcca taggctccgc cccccctgacg agcatcac
51 aaatcgacgc ggtggcgaaa cccgacagga ctataaagat accaggcg
```

JURASSIC PARK



- Download sequence from website

```
>DinoDNA  "Dinosaur DNA"
GAATTCCGGAAGCGAGCAAGAGATAAGTCCTGGCATCAGATACAGTTGGAGATAAGGACG
GACGTGTGGCAGCTCCCGCAGAGGATTCACTGGAAGTGCATTACCTATCCCATGGGAGCC
ATGGAGTTCGTGGCGCTGGGGGGGCCGGATGCGGGCTCCCCCACTCCGTTCCCTGATGAA
GCCGGAGCCTTCCTGGGGCTGGGGGGGGCGAGAGGACGGAGGCGGGGGGGCTGCTGGCC
TCCTACCCCCCCTCAGGCCGCGTGTCCCTGGTGCCGTGGGCAGACACGGGTACTTTGGGG
```

JURASSIC PARK

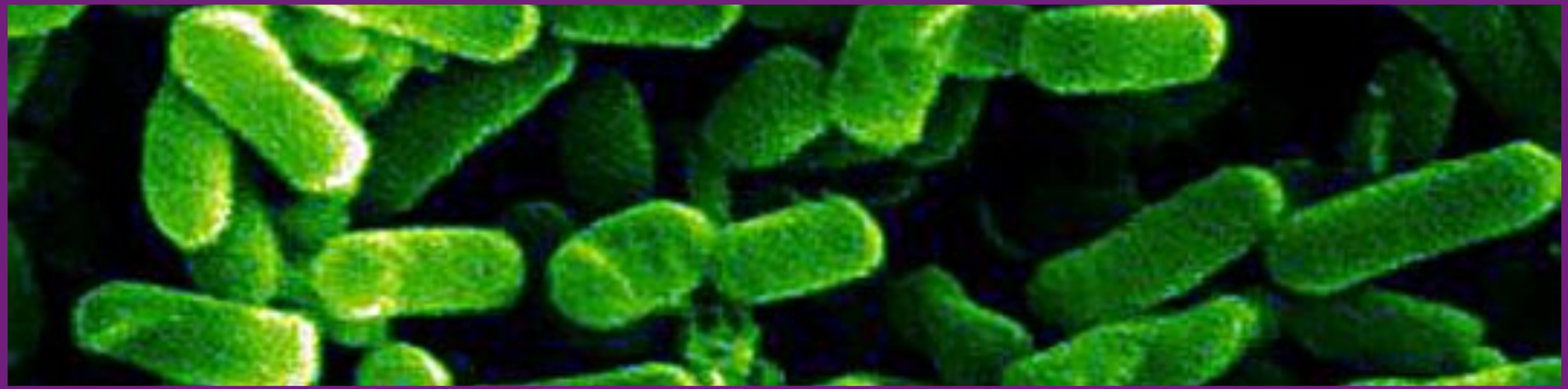
RANKED MOST SIMILAR TO LEAST

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Escherichia coli-Bacteroides shuttle vector pVAL-1 DNA, complete sequence	435	1415	98%	1e-117	88%	AB775653.1
Cloning vector pACYC184-ara-xylE, complete sequence	435	1042	97%	1e-117	88%	KJ170897.1
Cloning vector pFTC2, complete sequence	435	1367	93%	1e-117	88%	KF532952.1
Cloning vector pRMTn-Tc DNA, complete sequence	435	724	49%	1e-117	88%	AB777650.1
Cloning vector pRMR6K-Tc DNA, complete sequence	435	724	49%	1e-117	88%	AB777647.1
Cloning vector pLW18, complete sequence	435	1431	98%	1e-117	88%	KC820641.1
Cloning vector pLW17, complete sequence	435	1431	98%	1e-117	88%	KC820640.1
Cloning vector pBR939b, complete sequence	435	1431	98%	1e-117	88%	JQ394798.1
Cloning vector pMAT3, complete sequence	435	1317	93%	1e-117	88%	JN940924.1
Cloning vector pNG10A, complete sequence	435	1430	98%	1e-117	88%	JN940923.1
Cloning vector pGSC07, complete sequence	435	1258	86%	1e-117	88%	EU593760.1
Cloning vector pGSC03, complete sequence	435	1258	86%	1e-117	88%	EU593755.1
Cloning vector pWH1274, complete sequence	435	1431	98%	1e-117	88%	JN381160.1
Expression vector LIC-pDEST-LC2, complete sequence	435	1431	98%	1e-117	88%	JF327847.1
Expression vector LIC-pDEST-LC1, complete sequence	435	1431	98%	1e-117	88%	JF327846.1

JURASSIC PARK

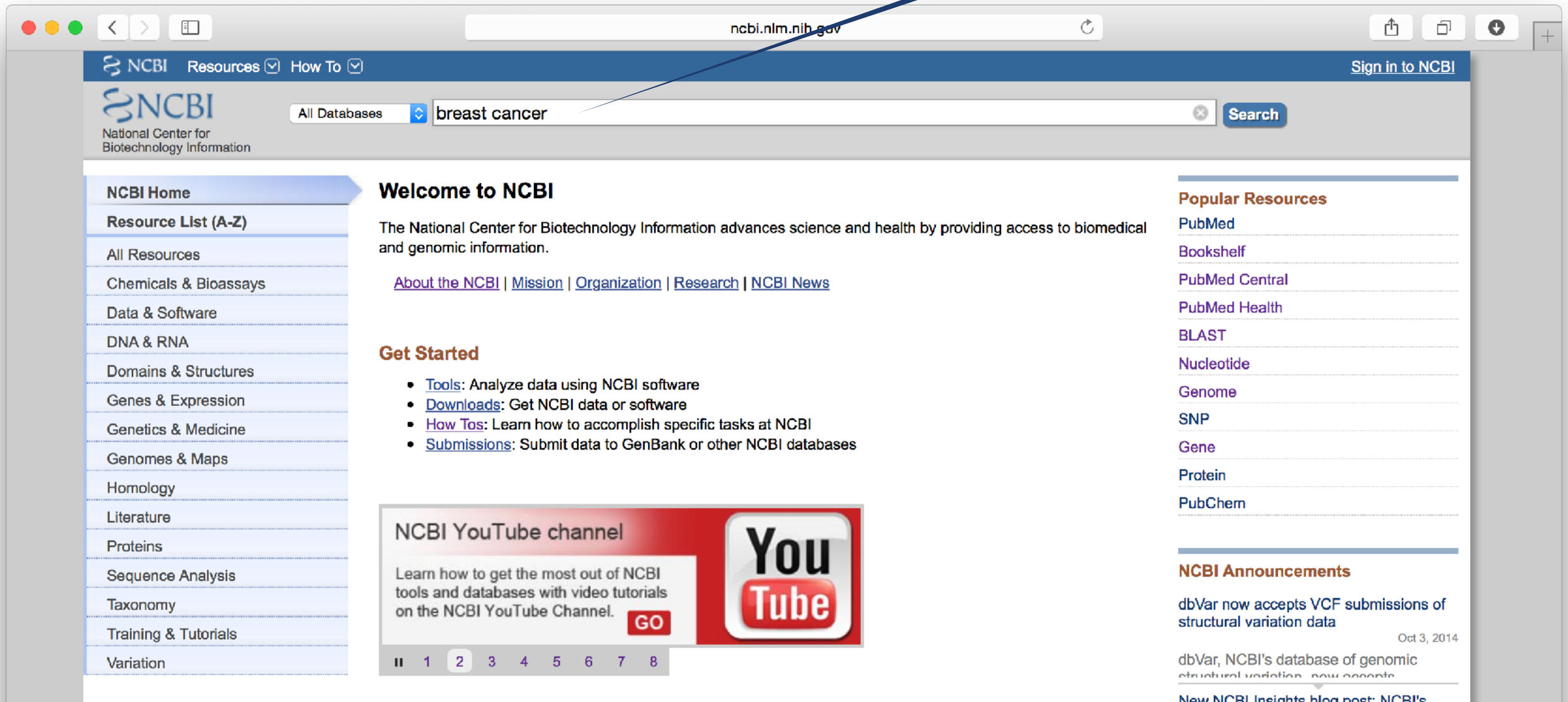


- E. coli
- Hollywood lied to us :(

BRCA1

BRCA1

"BREAST CANCER" SEARCH IN ENTREZ



BRCA1

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

GENE

Search NCBI databases

Help

breast cancer Search

About 5,480,244 search results for "breast cancer"

Breast cancer treatment - An overview from the US National **Cancer** Institute (NCI) - Patient version
Peer-reviewed, evidence-based summaries from the NCI's Physician Data Query (PDQ) database.
Health professional version

Literature

Books	12,240	books and reports
MeSH	52	ontology used for PubMed indexing
NLM Catalog	4,516	books, journals and more in the NLM Collections
PubMed	287,567	scientific & medical abstracts/citations
PubMed Central	222,672	full-text journal articles

Health

ClinVar	7,068	human variations of clinical significance
dbGaP	9,796	genotype/phenotype interaction studies
GTR	373	genetic testing registry
MedGen	290	medical genetics literature and links

Genes

EST	149,377	expressed sequence tag sequences
Gene	13,169	collected information about gene loci
GEO DataSets	51,706	functional genomics studies
GEO Profiles	4,313,749	gene expression and molecular abundance profiles
HomoloGene	87	homologous gene sets for selected organisms
PopSet	50	sequence sets from phylogenetic and population studies
UniGene	340	clusters of expressed transcripts

Proteins

Conserved Domains	118	conserved protein domains

BRCA1

GENE

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

Gene

Gene breast cancer Search

Save search Advanced Help

Show additional filters

Gene sources
Genomic
Mitochondria
Organelles
Plasmids

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
Current only

Chromosome locations
Select ...

Clear all

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance

Send to:

Results: 1 to 20 of 13169

<< First < Prev Page 1 of 659 Next > Last >>

Name/Gene ID	Description	Location	Aliases	MIM
FGFR2 ID: 2263	fibroblast growth factor receptor 2 [*Homo sapiens* (human)]	Chromosome 10, NC_000010.11 (121478330..121598458, complement)	BBDS, BEK, BFR-1, CD332, CEK3, CFD1, ECT1, JWS, K-SAM, KGFR, TK14, TK25	176943
ESR1 ID: 2099	estrogen receptor 1 [*Homo sapiens* (human)]	Chromosome 6, NC_000006.12 (151690496..152103274)	RP1-130E4.1, ER, ESR, ESRA, ESTRR, Era, NR3A1	133430
TOX3 ID: 27324	TOX high mobility group box family member 3 [*Homo sapiens* (human)]	Chromosome 16, NC_000016.10 (52437770..52547802, complement)	CAGF9, TNRC9	611416
CASC16 ID: 643714	cancer susceptibility candidate 16 (non-protein coding) [*Homo sapiens* (human)]	Chromosome 16, NC_000016.10 (52552087..52606975, complement)	RP11-297L17.1, LINC00918	
ZNF365 ID: 22891	zinc finger protein 365 [*Homo sapiens* (human)]	Chromosome 10, NC_000010.11 (62374157..62672011)	Su48, UAND, ZNF365	607818
TERT ID: 7015	telomerase reverse transcriptase [*Homo sapiens*	Chromosome 5, NC_000005.10	CMM9, DKCA2, DKCB4, EST2, PFBMFT1, TCS1, TP2, TRT,	187270

Hide sidebar >>

Filters: Manage Filters

Taxonomic Groups [List]

- eukaryotes *(10332)*
  - animals *(7488)*
    - chordates *(6716)*
      - vertebrates *(6678)*
      - more... *(38)*
    - more... *(772)*
  - fungi *(1426)*
    - ascomycetes *(1082)*
    - more... *(344)*
  - green plants *(875)*
    - land plants *(769)*
    - more... *(106)*
  - cellular slime molds *(74)*
  - kinetoplastids *(65)*
  - more... *(404)*
- bacteria *(2775)*
  - proteobacteria *(1139)*
  - firmicutes *(685)*
  - actinobacteria *(383)*
  - chlamydias *(117)*
  - CFB group bacteria *(96)*
  - mycoplasmas *(80)*
  - cyanobacteria *(77)*
  - more... *(198)*
- viruses *(36)*
- archaea *(26)*

Find related data

BRCA1

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

Search NCBI databases

Help

breast cancer Search

About 5,480,244 search results for "breast cancer"

Breast cancer treatment - An overview from the US National **Cancer** Institute
Patient version
Peer-reviewed, evidence-based summaries from the NCI's Physician Data Query (PDQ) database.
Health professional version

PUBMED

Literature

Books	12,240	books and reports
MeSH	52	ontology used for PubMed indexing
NLM Catalog	4,516	books, journals and more in the NLM Collections
PubMed	287,567	scientific & medical abstracts/citations
PubMed Central	222,672	full-text journal articles

Health

ClinVar	7,068	human variations of clinical significance
dbGaP	9,796	genotype/phenotype interaction studies
GTR	373	genetic testing registry
MedGen	290	medical genetics literature and links

Genes

EST	149,377	expressed sequence tag sequences
Gene	13,169	collected information about gene loci
GEO DataSets	51,706	functional genomics studies
GEO Profiles	4,313,749	gene expression and molecular abundance profiles
HomoloGene	87	homologous gene sets for selected organisms
PopSet	50	sequence sets from phylogenetic and population studies
UniGene	340	clusters of expressed transcripts

Proteins

Conserved Domains	118	conserved protein domains

BRCA1

ncbi.nlm.nih.gov

Irene L. Andrulis, Andrew K. Godwin, Drakoulis Yannoukakos, David E. Goldgar, Trinidad Caldes, Heli Nevanlinna, Ana Osorio, Mary Beth Terry, Mary B. Daly, Elizabeth J. van Rensburg, Ute Hamann, Susan J. Ramus, Amanda Ewart Toland, Maria A. Caligo, Olufunmilayo I. Olopade, Nadine Tung, Kathleen Claes, Mary S. Beattie, Melissa C. Southey, Evgeny N. Imyanitov, Marc Tischkowitz, Ramunas Janavicius, Esther M. John, Ava Kwong, Orland Diez, Judith Balmaña, Rosa B. Barkardottir, Banu K. Arun, Gad Rennert, Soo-Hwang Teo, Patricia A. Ganz, Ian Campbell, Annemarie H. van der Hout, Carolien H. M. van Deurzen, Caroline Seynaeve, Encarna B. Gómez Garcia, Flora E. van Leeuwen, Hanne E. J. Meijers-Heijboer, Johannes J. P. Gille, Margreet G. E. M. Ausems, Marinus J. Blok, Marjolijn J. L. Ligtenberg, Matti A. Rookus, Peter Devilee, Senno Verhoef, Theo A. M. van Os, Juul T. Wijnen, HEBON, EMBRACE, Debra Frost, Steve Ellis, Elena Fineberg, Radka Platte, D. Gareth Evans, Louise Izatt, Rosalind A. Eeles, Julian Adlard, Diana M. Eccles, Jackie Cook, Carole Brewer, Fiona Douglas, Shirley Hodgson, Patrick J. Morrison, Lucy E. Side, Alan Donaldson, Catherine Houghton, Mark T. Rogers, Huw Dorkins, Jacqueline Eason, Helen Gregory, Emma McCann, Alex Murray, Alain Calender, Agnès Hardouin, Pascaline Berthet, Capucine Delnatte, Catherine Nogues, Christine Lasset, Claude Houdayer, Dominique Leroux, Etienne Rouleau, Fabienne Prieur, Francesca Damiola, Hagay Sobol, Isabelle Coupier, Laurence Venat-Bouvet, Laurent Castera, Marion Gauthier-Villars, Mélanie Léoné, Pascal Pujol, Sylvie Mazoyer, Yves-Jean Bignon, GEMO Study Collaborators, Elżbieta Złowocka-Perłowska, Jacek Gronwald, Jan Lubinski, Katarzyna Durda, Katarzyna Jaworska, Tomasz Huzarski, Amanda B. Spurdle, Alessandra Viel, Bernard Peissel, Bernardo Bonanni, Giulia Melloni, Laura Ottini, Laura Papi, Liliana Varesco, Maria Grazia Tibiletti, Paolo Peterlongo, Sara Volorio, Siranoush Manoukian, Valeria Pensotti, Norbert Arnold, Christoph Engel, Helmut Deissler, Dorothea Gadzicki, Andrea Gehrig, Karin Kast, Kerstin Rhiem, Alfons Meindl, Dieter Niederacher, Nina Ditsch, Hansjoerg Plendl, Sabine Preisler-Adams, Stefanie Engert, Christian Sutter, Raymonda Varon-Mateeva, Barbara Wappenschmidt, Bernhard H. F. Weber, Brita Arver, Marie Stenmark-Askmalm, Niklas Loman, Richard Rosenquist, Zakaria Einbeigi, Katherine L. Nathanson, Timothy R. Rebbeck, Stephanie V. Blank, David E. Cohn, Gustavo C. Rodriguez, Laurie Small, Michael Friedlander, Victoria L. Bae-Jump, Anneliese Fink-Retter, Christine Rappaport, Daphne Gschwantler-Kaulich, Georg Pfeiler, Muy-Kheng Tea, Noralane M. Lindor, Bella Kaufman, Shani Shimon Paluch, Yael Laitman, Anne-Bine Skytte, Anne-Marie Gerdes, Inge Sokilde Pedersen, Sanne Traasdahl Moeller, Torben A. Kruse, Uffe Birk Jensen, Joseph Vijai, Kara Sarrel, Mark Robson, Noah Kauff, Anna Marie Mulligan, Gord Glendon, Hilmi Ozcelik, Bent Ejlertsen, Finn C. Nielsen, Lars Jønson, Mette K. Andersen, Yuan Chun Ding, Linda Steele, Lenka Foretova, Alex Teulé, Conxi Lazaro, Joan Brunet, Miquel Angel Pujana, Phuong L. Mai, Jennifer T. Loud, Christine Walsh, Jenny Lester, Sandra Orsulic, Steven A. Narod, Josef Herzog, Sharon R. Sand, Silvia Tognazzo, Simona Agata, Tibor Vaszko, Joellen Weaver, Alexandra V. Stavropoulou, Saundra S. Buys, Atocha Romero, Miguel de la Hoya, Kristiina Aittomäki, Taru A. Muranen, Mercedes Duran, Wendy K. Chung, Adriana Lasa, Cecilia M. Dorfling, Alexander Miron, BCFR, Javier Benitez, Leigha Senter, Dezheng Huo, Salina B. Chan, Anna P. Sokolenko, Jocelyne Chiquette, Laima Tihomirova, Tara M. Friebel, Bjarni A. Agnarsson, Karen H. Lu, Flavio Lejbkowicz, Paul A. James, Per Hall, Alison M. Dunning, Daniel Tessier, Julie Cunningham, Susan L. Slager, Chen Wang, Steven Hart, Kristen Stevens, Jacques Simard, Tomi Pastinen, Vernon S. Pankratz, Kenneth Offit, Douglas F. Easton, Georgia Chenevix-Trench, Antonis C. Antoniou, on behalf of CIMBA

PLoS Genet. 2013 March; 9(3): e1003212. Published online 2013 March 27. doi: 10.1371/journal.pgen.1003212

PMCID: PMC3609646

Article PubReader PDF–2.8M Supplementary Material

PUBMED

10. Common **breast cancer** susceptibility alleles are associated with tumour subtypes in *BRCA1* and *BRCA2* mutation carriers: results from the Consortium of Investigators of Modifiers of *BRCA1/2*

Anna Marie Mulligan, Fergus J Couch, Daniel Barrowdale, Susan M Domchek, Diana Eccles, Heli Nevanlinna, Susan J Ramus, Mark Robson, Mark Sherman, Amanda B Spurdle, Barbara Wappenschmidt, Andrew Lee, Lesley McGuffog, Sue Healey, Olga M Sinilnikova, Ramunas Janavicius, Thomas vO Hansen, Finn C Nielsen, Bent Ejlertsen, Ana Osorio, Iván Muñoz-Repeto, Mercedes Durán, Javier Godino, Maroulio Pertesi, Javier Benítez, Paolo Peterlongo, Siranoush Manoukian, Bernard Peissel, Daniela Zaffaroni,

BRCA1

ncbi.nlm.nih.gov

BioMed Central The Open Access Publisher | Breast Cancer RESEARCH

this article | search | submit a manuscript | register

Breast Cancer Res. 2011; 13(6): R110. PMCID: PMC3326552
Published online Nov 2, 2011. doi: 10.1186/bcr3052

Common breast cancer susceptibility alleles are associated with tumour subtypes in *BRCA1* and *BRCA2* mutation carriers: results from the Consortium of Investigators of Modifiers of *BRCA1/2*

Anna Marie Mulligan,[1] Fergus J Couch,[2] Daniel Barrowdale,[3] Susan M Domchek,[4] Diana Eccles,[5] Heli Nevanlinna,[6] Susan J Ramus,[7] Mark Robson,[8] Mark Sherman,[9] Amanda B Spurdle,[10] Barbara Wappenschmidt,[11] Andrew Lee,[3] Lesley McGuffog,[3] Sue Healey,[10] Olga M Sinilnikova,[12,13] Ramunas Janavicius,[14] Thomas vO Hansen,[15] Finn C Nielsen,[15] Bent Ejlertsen,[16] Ana Osorio,[17] Iván Muñoz-Repeto,[17] Mercedes Durán,[18] Javier Godino,[19] Maroulio Pertesi,[20] Javier Benitez,[21] Paolo Peterlongo,[22] Sirancush Manoukian,[23] Bernard Peissel,[23] Daniela Zaffaroni,[23] Elisa Cattaneo,[23] Bernardo Bonanni,[24] Alessandra Viel,[25] Barbara Pasini,[26] Laura Papi,[27] Laura Ottini,[28] Antonella Savarese,[29] Loris Bernard,[30] Paolo Radice,[22] Ute Hamann,[31] Martijn Verneus,[32] Hanne EJ Meijers-Heijboer,[33] Juul Wijnen,[34] Encarna B Gómez García,[35] Marcel R Nelen,[36] C Marleen Kets,[37] Caroline Seynaeve,[33] Madeleine MA Tilanus-Linthorst,[39] Rob B van der Luijt,[40] Theo van Os,[41] Matti Rookus,[32] Debra Frost,[3] J Louise Jones,[42] D Gareth Evans,[43] Fiona Lalloo,[43] Ros Eeles,[44] Louise Izatt,[45] Julian Adlard,[46] Rosemarie Davidson,[47] Jackie Cook,[48] Alan Donaldson,[49] Huw Dorkins,[50] Helen Gregory,[51] Jacqueline Eason,[52] Catherine Houghton,[53] Julian Barwell,[54] Lucy E Side,[55] Emma McCann,[56] Alex Murray,[57] Susan Peock,[3] Andrew K Godwin,[58] Rita K Schmutzler,[11] Kerstin Rhiem,[11] Christoph Engel,[59] Alfons Meindl,[60] Ina Rueh,[61] Norbert Arnold,[62] Dieter Niederacher,[63] Christian Sutter,[64] Helmut Deissler,[65] Dorothea Gadzicki,[66] Karin Kast,[67] Sabine Preisler-Adams,[68] Raymonda Varon-Mateeva,[69] Ines Schoenbuchner,[70] Britta Fiebig,[71] Wolfram Heinritz,[72] Dieter Schäfer,[73] Heidrun Gevensleben,[74] Virginie Caux-Moncoutier,[75] Marion Fassy-Colcombet,[75] François Cornelis,[76] Sylvie Mazoyer,[13] Mélanie Léoné,[12] Nadia Boutry-Kryza,[12] Agnès Hardouin,[77] Pascaline Berthet,[77] Danièle Muller,[78] Jean-Pierre Fricker,[78] Isabelle Mortemousque,[79] Pascal Pujol,[80] Isabelle Coupier,[80] Marine Lebrun,[81] Caroline Kientz,[81] Michel Longy,[82] Nicolas Sevenet,[82] Dominique Stoppa-Lyonnet,[75,83] Claudine Isaacs,[84] Trinidad Caldes,[85] Miguel de la Hoya,[85] Tuomas Heikkinen,[6] Kristiina Aittomäki,[86] Ignacio Blanco,[87] Conxi Lazaro,[87] Rosa B Barkardottir,[88] Penny Soucy,[89] Martine Dumont,[89] Jacques Simard,[90] Marco Montagna,[91] Silvia Tognazzo,[91] Emma D'Andrea,[92] Stephen Fox,[93] Max Yan,[94] Tim Rebbeck,[95] Olufunmilayo I Olopade,[96] Jeffrey N Weitzel,[97] Henry T Lynch,[98] Patricia A Ganz,[99] Gail E Tomlinson,[100] Xianshu Wang,[2] Zachary Fredericksen,[101] Vernon S Pankratz,[101] Noralane M Lindor,[102] Csilla Szabo,[103] Kenneth Offit,[8] Rita Sakr,[8] Mia Gaudet,[104] Jasmine Bhatia,[8] Noah Kauff,[8] Christian F Singer,[105] Muy-Kheng Tea,[105] Daphne Gschwantler-Kaulich,[105] Anneliese Fink-Retter,[105] Phuong L Mai,[106] Mark H Greene,[106] Evgeny Imyanitov,[107] Frances P O'Malley,[108] Hilmi Ozcelik,[109] Gordon Glendon,[110] Amanda E Toland,[111] Anne-Marie Gerdes,[112] Mads Thomassen,[113] Torben A Kruse,[113] Uffe Birk Jensen,[114] Anne-Bine Skytte,[115] Maria A Caligo,[116] Maria Soller,[117] Karin Henriksson,[118] von Anna Wachenfeldt,[119] Brita Arver,[119] Marie Stenmark-Askmalm,[120] Per Karlsson,[121] Yuan Chun Ding,[122] Susan L Neuhausen,[122] Mary Beattie,[123] Paul DP Pharoah,[124] Kirsten B Moysich,[125] Katherine L Nathanson,[4] Beth Y Karlan,[126] Jenny Gross,[126] Esther M John,[127] Mary B Daly,[128] Saundra M Buys,[129] Melissa C Southey,[130] John L Hopper,[131] Mary Beth Terry,[132] Wendy Chung,[132] Alexander F Miron,[133] David Goldgar,[134] Georgia Chenevix-Trench,[10] Douglas F Easton,[3] Irene L Andrulis,[135] and Antonis C Antoniou[3], Breast Cancer Family Registry

Breast Cancer Res

Formats:

Article | PubReader | ePub (beta) | PDF (471K)

Related citations in PubMed

Common alleles at 6q25.1 and 1p11.2 are associated with breast cancer risk for BRCA1 and BRCA2 mutatic [Hum Mol Genet. 2011]

Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carrie [Cancer Res. 2010]

Common variants at the 19p13.1 and ZNF365 loci are associated with ER subtypes of br [Cancer Epidemiol Biomarkers Prev. 2012]

Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. [Ann Oncol. 2011]

BRCA1/2 mutations and triple negative breast cancers. [Breast Dis. 2010]

See reviews...

See all...

Cited by other articles in PMC

A Phase II, Open-Label Study of Ramucirumab in Combination with Paclitaxel and Carboplat [Journal of Thoracic Oncology. 2014]

"Stop Ne(c)king around": How interactomics contributes to functionally charac [World Journal of Biological Chemistry. 2014...]

Regulation and roles of bicarbonate transporters in cancer [Frontiers in Physiology.]

The Divergence, Actions, Roles, and Relatives of Sodium-Coupled Bicarbonate Transporters [Physiological Reviews. 2013]

Metabotropic Glutamate Receptor 1 Expression and Its Polymorphic Variants Associate with Breast Cancer [PLoS ONE.]

...all...

Links

Cited in Books

EST

Gene

GEO Profiles

GENE

Display a menu

BRCA1

GENES FROM PAPER

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

Gene

Gene

Advanced

Help

Show additional filters

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Protein-coding

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
Current only

Chromosome locations
Select ...

Clear all

Show additional filters

Display Settings: Tabular, Sorted by Relevance

Send to:

Hide sidebar >>

Results: 2

Name/Gene ID	Description	Location	Aliases	MIM
BRCA2 ID: 675	breast cancer 2, early onset [*Homo sapiens* (human)]	Chromosome 13, NC_000013.11 (32315480..32399672)	RP11-298P3.4, BRCC2, BROVCA2, FACD, FAD, FAD1, FANCD, FANCD1, GLM3, PNCA2	600185
BRCA1 ID: 672	breast cancer 1, early onset [*Homo sapiens* (human)]	Chromosome 17, NC_000017.11 (43044295..43125483, complement)	BRCAI, BRCC1, BROVCA1, IRIS, PNCA4, PPP1R53, PSCP, RNF53	113705

Filters: Manage Filters

Find related data

Database: Select

Find items

Recent activity

Turn Off Clear

breast cancer (13169) Gene

BRCA1 breast cancer 1, early onset [Homo sapiens] Gene

breast cancer (222672) PMC

Gene Links for PMC (Select 3326552) (2) Gene

Common breast cancer susceptibility alleles are associated with tumour subtypes ... PMC

See more...

BRCA1

GENES FROM PAPER

Display Settings: Full Report

Hide sidebar >>

BRCA1 breast cancer 1, early onset [*Homo sapiens* (human)]

Gene ID: 672, updated on 5-Oct-2014

Summary

Official Symbol BRCA1 provided by HGNC
Official Full Name breast cancer 1, early onset provided by HGNC
Primary source HGNC:HGNC:1100
See related Ensembl:ENSG00000012048; HPRD:00218; MIM:113705; Vega:OTTHUMG00000157426
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as IRIS; PSCP; BRCAI; BRCC1; PNCA4; RNF53; BROVCA1; PPP1R53
Summary This gene encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. The encoded protein combines with other tumor suppressors, DNA damage sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC). This gene product associates with RNA polymerase II, and through the C-terminal domain, also interacts with histone deacetylase complexes. This protein thus plays a role in transcription, DNA repair of double-stranded breaks, and recombination. Mutations in this gene are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers. Alternative splicing plays a role in modulating the subcellular localization and physiological function of this gene. Many alternatively spliced transcript variants, some of which are disease-associated mutations, have been described for this gene, but the full-length natures of only some of these variants has been described. A related pseudogene, which is also located on chromosome 17, has been identified. [provided by RefSeq, May 2009]

Genomic context

Location: 17q21
Exon count: 24

See BRCA1 in Epigenomics, MapViewer

Table of contents

Summary
Genomic context
Genomic regions, transcripts, and products
Bibliography
Phenotypes
Variation
HIV-1 interactions
Pathways from BioSystems
Interactions
General gene information
Markers, Related pseudogene(s), Potential readthrough, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links
Locus-specific Databases

Related information

Order cDNA clone
3D structures
BioAssay
BioAssay by Target (List)
BioAssay by Target (Summary)

BRCA1

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	17	NC_000017.11 (43044295..43125483, complem
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	17	NC_000017.10 (41196312..41277500, complem

Chromosome 17 - NC_000017.11

[43014605 ▶ [43145255 ▶

VAT1 ◀ BRCA1 ◀ HMGN1P29 ▶
RND2 ▶ RPL21P4 ▶ NBR2 ▶
LOC101926920 ◀

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000017.11 chromosome 17 reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

NC_000017.11: 43M..43M (106Kbp) C ▾ Find: - + Tools ▾ Configure

40 K | 43,130 K | 43,120 K | 43,110 K | 43,100 K | 43,090 K | 43,080 K | 43,070 K | 43,060 K | 43,050 K | 43,040 K

Genes, NCBI Homo sapiens Annotation Release 106

BRCA1
NM_007294.3
NP_009225.1
NM_007300.3
NP_009231.2
NM_007297.3
NP_009228.2
NM_007299.3
NP_009230.2
XM_006722029.1
XP_006722092.1
XM_006722031.1
XP_006722094.1
XM_006722035.1
XP_006722098.1
XM_006722039.1
XP_006722102.1

NM_007294.3
mRNA: NM_007294.3
Title: mRNA-breast cancer 1, early onset, transcript variant 1
Location: complement(43,044,295..43,125,483)
[*Length*]
Span: 81,189
Placed: 7,207
Product: 7,224

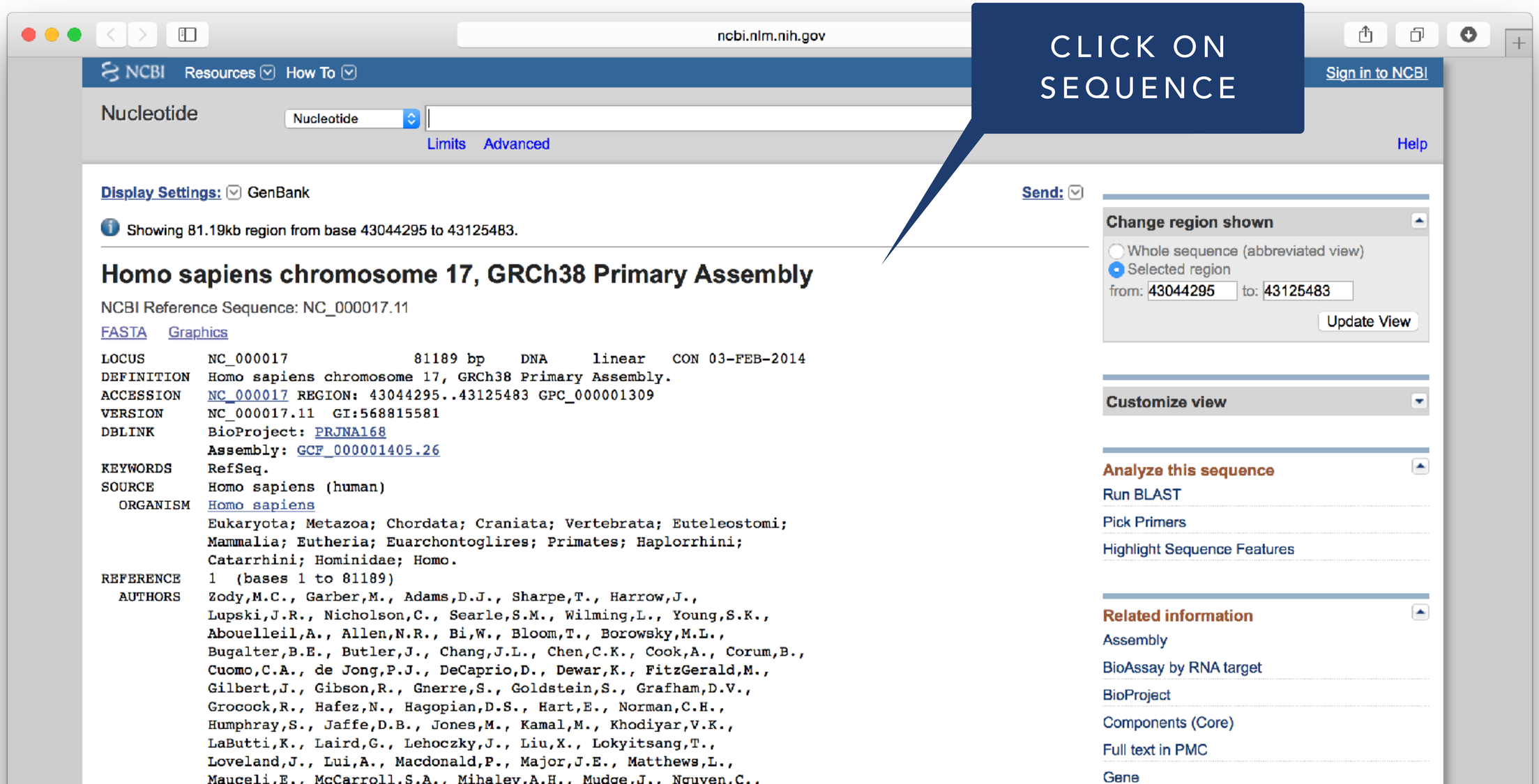
Download: NM_007294.3

Links & Tools
View GeneID: 672 (BRCA1)

CLICK ON SEQUENCE, SHOW GENBANK

BioAssay, by Gene target
ClinVar
Conserved Domains
dbVar
EST
Full text in PMC
Full text in PMC_nucleotide
Gene neighbors
Genome
GEO Profiles
GTR
HomoloGene
Map Viewer
MedGen
Nucleotide
OMIM
Probe
Protein
PubChem Compound
PubChem Substance

BRCA1



CLICK ON SEQUENCE

BRCA1

```
                /chromosome="17"
                /map="17q21"
gene            1..7224
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
                /note="breast cancer 1, early onset"
                /db_xref="GeneID:672"
                /db_xref="HGNC:HGNC:1100"
                /db_xref="HPRD:00218"
                /db_xref="MIM:113705"
exon            1..213
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
                /inference="alignment:Splign:1.39.8"
misc_feature    93
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
                /note="alternative transcription start site"
STS             184..1651
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
                /db_xref="UniSTS:482836"
misc_feature    209..211
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
                /note="upstream in-frame stop codon"
exon            214..312
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
                /inference="alignment:Splign:1.39.8"
CDS             233..5824
                /gene="BRCA1"
                /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                PPP1R53; PSCP; RNF53"
```

CLICK ON SEQUENCE

Gene

OMIM

Protein

PubMed

PubMed (RefSeq)

PubMed (Weighted)

SNP

Taxonomy

UniGene

GEO Profiles

Recent activity

Turn Off Clear

Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 1, mRNA Nucleotide

BRCA1 breast cancer 1, early onset [Homo sapiens] Gene

Homo sapiens breast cancer 1, early onset (BRCA1), RefSeqGene (LRG_292) (Nucleotide

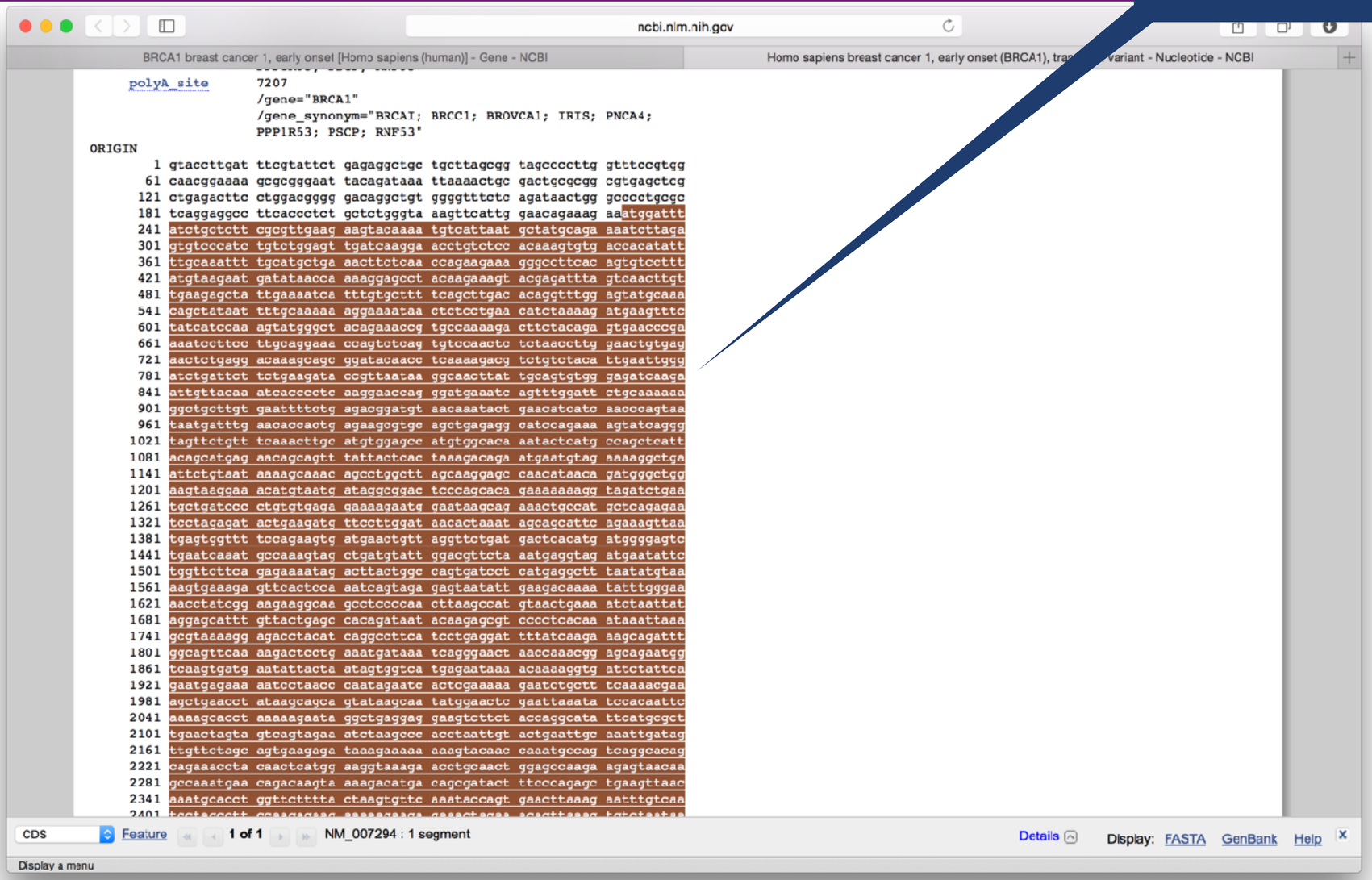
Gene Links for PMC (Select 3326552) (2) Gene

Common breast cancer susceptibility alleles are associated with tumour subtypes ... PMC

See more...

BRCA1

CLICK ON SEQUENCE



BRCA1

```
 121 ctgagacttc ctggacgggg gacaggctgt ggggtttctc agataactgg gccccctgcgc
 181 tcaggaggcc ttcaccctct gctctgggta aagttcattg gaacagaaag aaatggattt
 241 atctgctctt cgcgttgaag aagtacaaaa tgtcattaat gctatgcaga aaatcttaga
 301 gtgtcccatc tgtctggagt tgatcaagga acctgtctcc acaaagtgtg accacatatt
 361 ttgcaaattt tgcatgctga aacttctcaa ccagaagaaa gggccttcac agtgtccttt
 421 atgtaagaat gatataacca aaaggagcct acaagaaagt acgagattta gtcaacttgt
 481 tgaagagcta ttgaaaatca tttgtgcttt tcagcttgac acaggtttgg agtatgcaaa
 541 cagctataat tttgcaaaaa aggaaaataa ctctcctgaa catctaaaag atgaagtttc
 601 tatcatccaa agtatgggct acagaaaccg tgccaaaaga cttctacaga gtgaacccga
 661 aaatccttcc ttgcaggaaa ccagtctcag tgtccaactc tctaaccttg gaactgtgag
 721 aactctgagg acaaagcagc ggatacaacc tcaaaagacg tctgtctaca ttgaattggg
 781 atctgattct tctgaagata ccgttaataa ggcaacttat tgcagtgtgg gagatcaaga
 841 attgttacaa atcacccctc aaggaaccag ggatgaaatc agtttggatt ctgcaaaaaa
 901 ggctgcttgt gaattttctg agacggatgt aacaaatact gaacatcatc aacccagtaa
 961 taatgatttg aacaccactg agaagcgtgc agctgagagg catccagaaa agtatcaggg
1021 tagttctgtt tcaaacttgc atgtggagcc atgtggcaca aatactcatg ccagctcatt
1081 acagcatgag aacagcagtt tattactcac taaagacaga atgaatgtag aaaaggctga
1141 attctgtaat aaaagcaaac agcctggctt agcaaggagc caacataaca gatgggctgg
1201 aagtaaggaa acatgtaatg ataggcggac tcccagcaca gaaaaaaagg tagatctgaa
1261 tgctgatccc ctgtgtgaga gaaaagaatg gaataagcag aaactgccat gctcagagaa
1321 tcctagagat actgaagatg ttccttggat aacactaaat agcagcattc agaaagttaa
1381 tgagtggttt tccagaagtg atgaactgtt aggttctgat gactcacatg atggggagtc
1441 tgaatcaaat gccaaagtag ctgatgtatt ggacgttcta aatgaggtag atgaatattc
1501 tggttcttca gagaaaatag acttactggc cagtgatcct catgaggctt taatatgtaa
1561 aagtgaaaga gttcactcca aatcagtaga gagtaatatt gaagacaaaa tatttgggaa
1621 aacctatcgg aagaaggcaa gcctccccaa cttaagccat gtaactgaaa atctaattat
1681 aggagcattt gttactgagc cacagataat acaagagcgt cccctcacaa ataaattaaa
1741 gcgtaaaagg agacctacat caggccttca tcctgaggat tttatcaaga aagcagattt
1801 ggcagttcaa aagactcctg aaatgataaa tcagggaact aaccaaacgg agcagaatgg
1861 tcaagtgatg aatattacta atagtggtca tgagaataaa acaaaaggtg attctattca
1921 gaatgagaaa aatcctaacc caatagaatc actcgaaaaa gaatctgctt tcaaaacgaa
1981 agctgaacct ataagcagca gtataagcaa tatggaactc gaattaaata tccacaattc
2041 aaaagcacct aaaaagaata ggctgaggag gaagtcttct accaggcata ttcatgcgct
2101 tgaactagta gtcagtagaa atctaagccc acctaattgt actgaattgc aaattgatag
2161 ttgttctagc agtgaagaga taaagaaaaa aaagtacaac caaatgccag tcaggcacag
2221 cagaaaccta caactcatgg aaggtaaaga acctgcaact ggagccaaga agagtaacaa
2281 gccaaatgaa cagacaagta aaagacatga cagcgatact ttcccagagc tgaagttaac
```

CLICK ON FEATURES TO SEE DETAILS

```
233..235
/gene="BRCA1"
/gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
PPP1R53; PSCP; RNF53"
/experiment="experimental evidence, no additional details
recorded"
/note="N-acetylmethionine; propagated from
UniProtKB/Swiss-Prot (P38398.2); acetylation site"
```

misc_feature Feature 3 of 35 NM_007294 : 1 segment Details Display: FASTA GenBank Help

BRCA1

SAVE SEQUENCE IN FASTA

Display Settings: GenBank

Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 1, mRNA

NCBI Reference Sequence: NM_007294.3

FASTA Graphics

Go to:

```
LOCUS       NM_007294               7224 bp    mRNA    linear   PRI 03-MAY-2014
DEFINITION  Homo sapiens breast cancer 1, early onset (BRCA1), transcript
            variant 1, mRNA.
ACCESSION   NM_007294
VERSION     NM_007294.3  GI:237757283
KEYWORDS    RefSeq.
SOURCE      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 7224)
  AUTHORS   Orban TI and Olah E.
  TITLE     Emerging roles of BRCA1 alternative splicing
  JOURNAL   MP, Mol. Pathol. 56 (4), 191-197 (2003)
   PUBMED   12890739
  REMARK    Review article
REFERENCE   2  (bases 1 to 7224)
  AUTHORS   Orban TI and Olah E.
  TITLE     Expression profiles of BRCA1 splice variants in asynchronous and in
            G1/S synchronized tumor cell lines
  JOURNAL   Biochem. Biophys. Res. Commun. 280 (1), 32-38 (2001)
   PUBMED   11162473
REFERENCE   3  (bases 1 to 7224)
  AUTHORS   Paterson JW.
  TITLE     BRCA1: a review of structure and putative functions
  JOURNAL   Dis. Markers 13 (4), 261-274 (1998)
   PUBMED   9553742
  REMARK    Review article
```

Change region shown

Customize view

Basic Features
- (•) Default features
- ( ) Gene, RNA, and CDS features only

Features added by NCBI
- [ ] 2159 SNPs
- [x] 5 conserved domains

Display options
- [x] Show sequence
- [ ] Show reverse complement

Update View

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

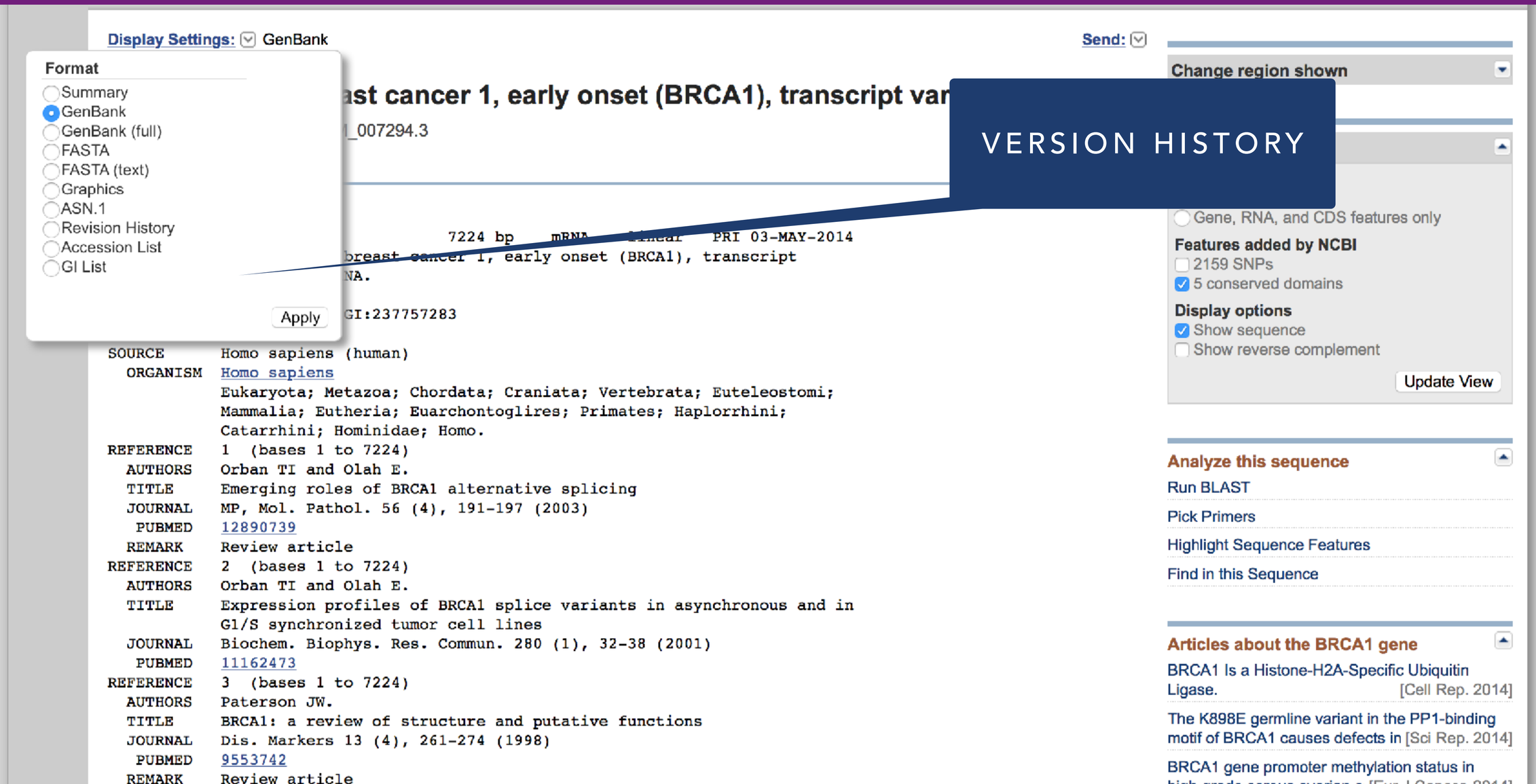
Articles about the BRCA1 gene

BRCA1 Is a Histone-H2A-Specific Ubiquitin Ligase. [Cell Rep. 2014]

The K898E germline variant in the PP1-binding motif of BRCA1 causes defects in [Sci Rep. 2014]

BRCA1 gene promoter methylation status in

BRCA1


Display Settings: GenBank
Send:
Format
Summary
GenBank
GenBank (full)
FASTA
FASTA (text)
Graphics
ASN.1
Revision History
Accession List
GI List
Apply
ast cancer 1, early onset (BRCA1), transcript var
_007294.3
VERSION HISTORY
7224 bp mRNA PRI 03-MAY-2014
breast cancer 1, early onset (BRCA1), transcript
NA.
GI:237757283
SOURCE 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 7224)
AUTHORS Orban TI and Olah E.
TITLE Emerging roles of BRCA1 alternative splicing
JOURNAL MP, Mol. Pathol. 56 (4), 191-197 (2003)
PUBMED 12890739
REMARK Review article
REFERENCE 2 (bases 1 to 7224)
AUTHORS Orban TI and Olah E.
TITLE Expression profiles of BRCA1 splice variants in asynchronous and in
G1/S synchronized tumor cell lines
JOURNAL Biochem. Biophys. Res. Commun. 280 (1), 32-38 (2001)
PUBMED 11162473
REFERENCE 3 (bases 1 to 7224)
AUTHORS Paterson JW.
TITLE BRCA1: a review of structure and putative functions
JOURNAL Dis. Markers 13 (4), 261-274 (1998)
PUBMED 9553742
REMARK Review article
Change region shown
Gene, RNA, and CDS features only
Features added by NCBI
2159 SNPs
5 conserved domains
Display options
Show sequence
Show reverse complement
Update View
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Articles about the BRCA1 gene
BRCA1 Is a Histone-H2A-Specific Ubiquitin Ligase. [Cell Rep. 2014]
The K898E germline variant in the PP1-binding motif of BRCA1 causes defects in [Sci Rep. 2014]
BRCA1 gene promoter methylation status in

BRCA1

ncbi.nlm.nih.gov

```
            RNAseq introns              :: mixed/partial sample support
                                           ERS025081, ERS025082 [ECO:0000350]
            ##Evidence-Data-END##
            COMPLETENESS: full length.
PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
            1-143               AL701927.1         1-143
            144-5824            U14680.1           31-5711
            5825-6573           BC072418.1         2401-3149
            6574-7205           BU617173.1         19-650              c
            7206-7224           BU679389.1         1-19                c
FEATURES             Location/Qualifiers
     source          1..7224
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="17"
                     /map="17q21"
     gene            1..7224
                     /gene="BRCA1"
                     /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                     PPP1R53; PSCP; RNF53"
                     /note="breast cancer 1, early onset"
                     /db_xref="GeneID:672"
                     /db_xref="HGNC:HGNC:1100"
                     /db_xref="HPRD:00218"
                     /db_xref="MIM:113705"
     exon            1..213
                     /gene="BRCA1"
                     /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                     PPP1R53; PSCP; RNF53"
                     /inference="alignment:Splign:1.39.8"
     misc_feature    93
                     /gene="BRCA1"
                     /gene_synonym="BRCAI; BRCC1; BROVCA1; IRIS; PNCA4;
                     PPP1R53; PSCP; RNF53"
                     /note="alternative transcription start site"
     STS             184..1651
```

LINK TO PROTEIN

UCSC Genome Browser
[UCSC Genome Browser]

Related information

- Related Sequences
- BioSystems
- CCDS
- Components (Core)
- Components (EST)
- Full text in PMC
- Gene
- GeneView in dbSNP
- Map Viewer
- Master
- OMIM
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- SNP
- Taxonomy
- UniGene
- GEO Profiles

BRCA1

Display Settings: GenPept

Send to:

breast cancer type 1 susceptibility protein isoform 1 [Homo sapiens]

LINK TO PROTEIN

NCBI Reference Sequence: NP_009225.1

FASTA Graphics

Go to:

```
LOCUS       NP_009225                1863 aa            linear   PRI 03-MAY-2014
DEFINITION  breast cancer type 1 susceptibility protein isoform 1 [Homo
            sapiens].
ACCESSION   NP_009225
VERSION     NP_009225.1  GI:6552299
DBSOURCE    REFSEQ: accession NM_007294.3
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (sites)
  AUTHORS   Matsuoka,S., Ballif,B.A., Smogorzewska,A., McDonald,E.R. III,
            Hurov,K.E., Luo,J., Bakalarski,C.E., Zhao,Z., Solimini,N.,
            Lerenthal,Y., Shiloh,Y., Gygi,S.P. and Elledge,S.J.
  TITLE     ATM and ATR substrate analysis reveals extensive protein networks
            responsive to DNA damage
  JOURNAL   Science 316 (5828), 1160-1166 (2007)
   PUBMED   17525332
REFERENCE   2  (sites)
  AUTHORS   Olsen,J.V., Blagoev,B., Gnad,F., Macek,B., Kumar,C., Mortensen,P.
            and Mann,M.
  TITLE     Global, in vivo, and site-specific phosphorylation dynamics in
            signaling networks
  JOURNAL   Cell 127 (3), 635-648 (2006)
   PUBMED   17081983
REFERENCE   3  (sites)
  AUTHORS   Fabbro,M., Savage,K., Hobson,K., Deans,A.J., Powell,S.N.,
```

Analyze this sequence

- Run BLAST
- Identify Conserved Domains
- Highlight Sequence Features
- Find in this Sequence

Protein 3D Structure

Crystal Structure Of Brca1 Brct With Doubly Phosphorylated Abraxas

PDB: 4JLU
Source: Homo sapiens
Method: X-Ray Diffraction
Resolution: 3.5 Å

See all 23 structures...

Articles about the BRCA1 gene

- BRCA1 gene promoter methylation status in high-grade serous ovarian c [Eur J Cancer. 2014]
- Identification of a BRCA1-mRNA splicing complex required for efficient DN [Mol Cell. 2014]
- BRCA1 deficiency exacerbates estrogen-induced

BRCA1

```
   PUBMED   8840964
REFERENCE   21 (residues 1 to 1863)
  AUTHORS   Holt JT, Thompson ME, Szabo C, Robinson-Benion C, Artea
            MC and Jensen RA.
  TITLE     Growth retardation and tumour inhibition by BRCA1
  JOURNAL   Nat. Genet. 12 (3), 298-302 (1996)
   PUBMED   8589721
  REMARK    Erratum:[Nat Genet 1998 May;19(1):102]
REFERENCE   22 (residues 1 to 1863)
  AUTHORS   Xu CF, Brown MA, Chambers JA, Griffiths B, Nicolai H and Solomon E.
  TITLE     Distinct transcription start sites generate two forms of BRCA1 mRNA
  JOURNAL   Hum. Mol. Genet. 4 (12), 2259-2264 (1995)
   PUBMED   8634696
REFERENCE   23 (residues 1 to 1863)
  AUTHORS   Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K.
            Tavtigian S, Liu Q, Cochran C, Bennett LM and Ding
  CONSRTM   et al
  TITLE     A strong candidate for the breast and ovarian canc
            gene BRCA1
  JOURNAL   Science 266 (5182), 66-71 (1994)
   PUBMED   7545954
REFERENCE   24 (residues 1 to 1863)
  AUTHORS   Petrucelli,N., Daly,M.B. and Feldman,G.L.
  TITLE     BRCA1 and BRCA2 Hereditary Breast and Ovarian Cancer
  JOURNAL   (in) Pagon RA, Adam MP, Ardinger HH, Bird TD, Dolan CR, Fong CT,
            Smith RJH and Stephens K (Eds.);
            GENEREVIEWS(R);
            (1993)
   PUBMED   20301425
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AL701927.1, U14680.1,
            BC072418.1, BU617173.1 and BU679389.1.
            This sequence is a reference standard in the RefSeqGene project.

            Summary: This gene encodes a nuclear phosphoprotein that plays a
            role in maintaining genomic stability, and it also acts as a tumor
            suppressor. The encoded protein combines with other tumor
            suppressors, DNA damage sensors, and signal transducers to form a
            large multi-subunit protein complex known as the BRCA1-associated
            genome surveillance complex (BASC). This gene product associates
```

LINK TO OMIM

Online Mendelian Inheritance in Man (OMIM) records that refer to protein records in the current set.

- Conserved Domains (Full)
- Domain Relatives
- Encoding mRNA
- Full text in PMC
- Gene
- Gene Genotype
- GeneView in dbSNP
- Genome
- Map Viewer
- Nucleotide
- OMIM
- Protein (UniProtKB)
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Related Structures (List)
- Related Structures (Summary)
- SNP
- Structure
- Taxonomy
- UniGene

Recent activity

Turn Off Clear

- breast cancer type 1 susceptibility protein isoform 1 [Homo sapiens] Protein
- Homo sapiens breast cancer 1, early onset

BRCA1

Gene summaries Genetic tests Medical literature

2. *607585 - ATAXIA-TELANGIECTASIA MUTATED GENE; ATM
Cytogenetic locations: 11q22.3
OMIM: 607585
Gene summaries Genetic tests Medical literature

3. +604373 - CHECKPOINT KINASE 2, S. POMBE, HOMOLOG OF; CHEK2
BREAST AND COLORECTAL CANCER, SUSCEPTIBILITY TO, INCLUDED
Cytogenetic locations: 22q12.1
OMIM: 604373
Gene summaries Genetic tests Medical literature

4. #604370 - BREAST-OVARIAN CANCER, FAMILIAL, SUSCEPTIBILITY TO, 1; BROVCA1
BREAST CANCER, FAMILIAL, SUSCEPTIBILITY TO, 1, INCLUDED
Cytogenetic locations: 17q21.31
OMIM: 604370
Gene summaries Genetic tests Medical literature

5. *601215 - ATR GENE; ATR
Cytogenetic locations: 3q23
OMIM: 601215
Gene summaries Genetic tests Medical literature

6. #114480 - BREAST CANCER
BREAST CANCER, FAMILIAL MALE, INCLUDED
Cytogenetic locations: 2q35, 1pter-p36.13, 2q33.1, 1pter-p36.13, 22q12.1, 1pter-p36.13, 17q23.2, 1pter-p36.13, 17q23.2, 1pter-p36.13, 17q21.33, 1pter-p36.13, 17p13.1, 1pter-p36.13, 16q22.1, 1pter-p36.13, 16p12.2, 1pter-p36.13, 15q15.1, 12q35, 1pter-p36.13, 2q33.1, 1pter-p36.13, 22q12.1, 1pter-p36.13, 17q23.2, 1pter-p36.13, 17q23.2, 1pter-p36.13, 17q21.33, 1pter-p36.13, 17p13.1, 1pter-p36.13, 16q22.1, 1pter-p36.13, 16p12.2, 1pter-p36.13, 15q15.1,
OMIM: 114480
Gene summaries Genetic tests Medical literature

7. *113705 - BREAST CANCER 1 GENE; BRCA1
Cytogenetic locations: 17q21.31

BRCA1 ENTRY

Find related data
Database: Select
Find items

Recent activity
Turn Off Clear
OMIM Links for Nucleotide (Select 237757283) (7) OMIM
OMIM Links for Protein (Select 6552299) (6) OMIM
breast cancer type 1 susceptibility protein isoform 1 [Homo sapiens] Protein
Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 1, mRNA Nucleotide
BRCA1 breast cancer 1, early onset [Homo sapiens] Gene
See more...

BRCA1

omim.org

Home | About | Statistics | Downloads | Help | External Links | Terms of Use | Contact Us | MIMmatch ← NEW | Select Language

Search OMIM... Search

Advanced Search | Display Options

*113705

BREAST CANCER 1 GENE; BRCA1

HGNC Approved Gene Symbol: ***BRCA1***

Cytogenetic location: ***17q21.31*** ***Genomic coordinates (GRCh37):*** ***17:41,196,311-41,277,499*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Phenotype mapping key
17q21.31	{Breast-ovarian cancer, familial, 1}	604370	3
	{Pancreatic cancer, susceptibility to, 4}	614320	3

TEXT

Description

BRCA1 plays critical roles in DNA repair, cell cycle checkpoint control, and maintenance of genomic stability. BRCA1 forms several distinct complexes through association with different adaptor proteins, and each complex forms in a mutually exclusive manner (Wang et al., 2009).

Cloning and Expression

Table of Contents for *113705

- Title
- Gene-Phenotype Relationships
- Text
  - Description
  - Cloning and Expression
  - Gene Structure
  - Mapping
  - Gene Function
  - Molecular Genetics
  - Genotype/Phenotype Correlations
  - Evolution
  - Animal Model
- Allelic Variants
  - Table View
- See Also
- References
- Contributors
- Creation Date
- Edit History

External Links for Entry:

- Genome
- DNA
- Protein

SARS-CoV-2 (CORONAVIRUS)

CORONAVIRUS

- Virus
  - Severe acute respiratory syndrome coronavirus 2
  - SARS-CoV-2
- Coronavirus disease
  - 2019 novel coronavirus
  - COVID-19

Home / Emergencies / Diseases / Coronavirus disease 2019 / Technical guidance /
Naming the coronavirus disease (COVID-19) and the virus that causes it

Section navigation

Naming the coronavirus disease (COVID-19) and the virus that causes it

Official names have been announced for the virus responsible for COVID-19 (previously known as "2019 novel coronavirus") and the disease it causes. The official names are:

Disease

coronavirus disease
(COVID-19)

Virus

severe acute respiratory syndrome coronavirus 2
(SARS-CoV-2)

SIMILAR TO SARS FROM 2003

Why do the virus and the disease have different names?

Viruses, and the diseases they cause, often have different names. For example, HIV is the virus that causes AIDS. People often know the name of a disease, but not the name of the virus that causes it.

There are different processes, and purposes, for naming viruses and diseases.

CORONAVIRUS

GenBank ▾ Send to: ▾

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

GenBank: MN908947.3

FASTA Graphics

Go to:

```
LOCUS       MN908947                29903 bp ss-RNA     linear   VRL 18-MAR-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   MN908947
VERSION     MN908947.3
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae;
            Orthocoronavirinae; Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29903)
  AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.M.
            Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H.,
```

HTTPS://WWW.NCBI.NLM.NIH.GOV/NUCCORE/MN908947.3

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

NCBI Virus

Retrieve, view, and download S
CoV-2 coronavirus genomic an
sequences.

CORONAVIRUS

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Search
Advanced | Help

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.

The new PubMed site will become the default in mid-May.
Click here to try it now!
Frequently asked questions

Format: Abstract | Send to

Nature. 2020 Mar;579(7798):265-269. doi: 10.1038/s41586-020-2008-3. Epub 2020 Feb 3.

A new coronavirus associated with human respiratory disease in China.

Wu F[1], Zhao S[2], Yu B[3], Chen YM[1], Wang W[4], Song ZG[1], Hu Y[2], Tao ZW[2], Tian JH[3], Pei YY[1], Yuan ML[2], Zhang YL[1], Dai FH[1], Liu Y[1], Wang QM[1], Zheng JJ[1], Xu L[1], Holmes EC[1,5], Zhang YZ[6,7,

Full text links
nature
PMC FREE Full text

HTTPS://WWW.NCBI.NLM.NIH.GOV/PUBMED/32015508

CORONAVIRUS

nature

Search E-alert Submit Login

Article | Open Access | Published: 03 February 2020

A new coronavirus associated with human respiratory disease in China

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes & Yong-Zhen Zhang

Nature **579**, 265–269(2020) | Cite this article

Download PDF

Associated Content

Collection

Coronavirus

Sections Figures References

HTTPS://WWW.NATURE.COM/ARTICLES/S41586-020-2008-3

CORONAVIRUS

NCBI Resources How To Sign in to NCBI

Protein

Protein Search

Advanced Help

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.

Species
Viruses (10)
Customize ...

Source databases
Customize ...

Sequence length
Custom range...

Molecular

Summary 20 per page Sort by Default order Send to:

Filters: Manage Filters

Items: 10

1. ORF10 protein [Severe acute respiratory syndrome coronavirus 2]
38 aa protein
Accession: QHI42199.1 GI: 1798172433
Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

Analyze these sequences
Run BLAST
Align sequences with COBALT
Identify Conserved Domains with CD-Search

HTTPS://WWW.NCBI.NLM.NIH.GOV/PROTEIN?LINKNAME=PUBMED_PROTEIN&FROM_UID=32015508

CORONAVIRUS

```
                GQINDMILSLLSKGRLIIRENNRVVISSDVLVNN"
gene            21563..25384
                /gene="S"
CDS             21563..25384
                /gene="S"
                /note="structural protein"
                /codon_start=1
                /product="surface glycoprotein"
                /protein_id="QHD43416.1"
                /translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFR
                SSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR
                GWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVY
                SSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQ
                GFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL
                LKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITN
                LCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
                TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
                YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
                RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG
                RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAI
                HADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPR
                RARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTM
                YICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG
                GFNFSQILPDPSKPSKRSFIEDLL
                GLTVLPPLLTDEMIAQYTSALLA
```

HTTPS://WWW.NCBI.NLM.NIH.GOV/NUCCORE/MN908947.3

CORONAVIRUS

```
                GQINDMILSLLSKGRLIIRENNRVVISSDVLVNN"
gene            21563..25384
                /gene="S"
CDS             21563..25384
                /gene="S"
                /note="structural protein"
                /codon_start=1
                /product="surface glycoprotein"
                /protein_id="QHD43416.1"
                /translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFR
                SSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR
                GWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVY
                SSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQ
                GFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL
                LKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITN
                LCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
                TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
                YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
                RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG
                RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAI
                HADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPR
                RARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTM
                YICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG
                GFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFN
                GLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQN
                VLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA
```

HTTPS://WWW.NCBI.NLM.NIH.GOV/NUCCORE/MN908947.3

CORONAVIRUS

GenPept ▾ Send to: ▾

surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]

GenBank: QHD43416.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       QHD43416                1273 aa            linear   VRL 18-MAR-2020
DEFINITION  surface glycoprotein [Severe acute respiratory syndrome coronavirus
            2].
ACCESSION   QHD43416
VERSION     QHD43416.1
DBSOURCE    accession MN908947.3
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae;
            Orthocoronavirinae; Betacoronavirus; Sarbecovirus.
REFERENCE   1  (residues 1 to 1273)
  AUTHORS   Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.G., Hu Y.
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

NCBI Virus

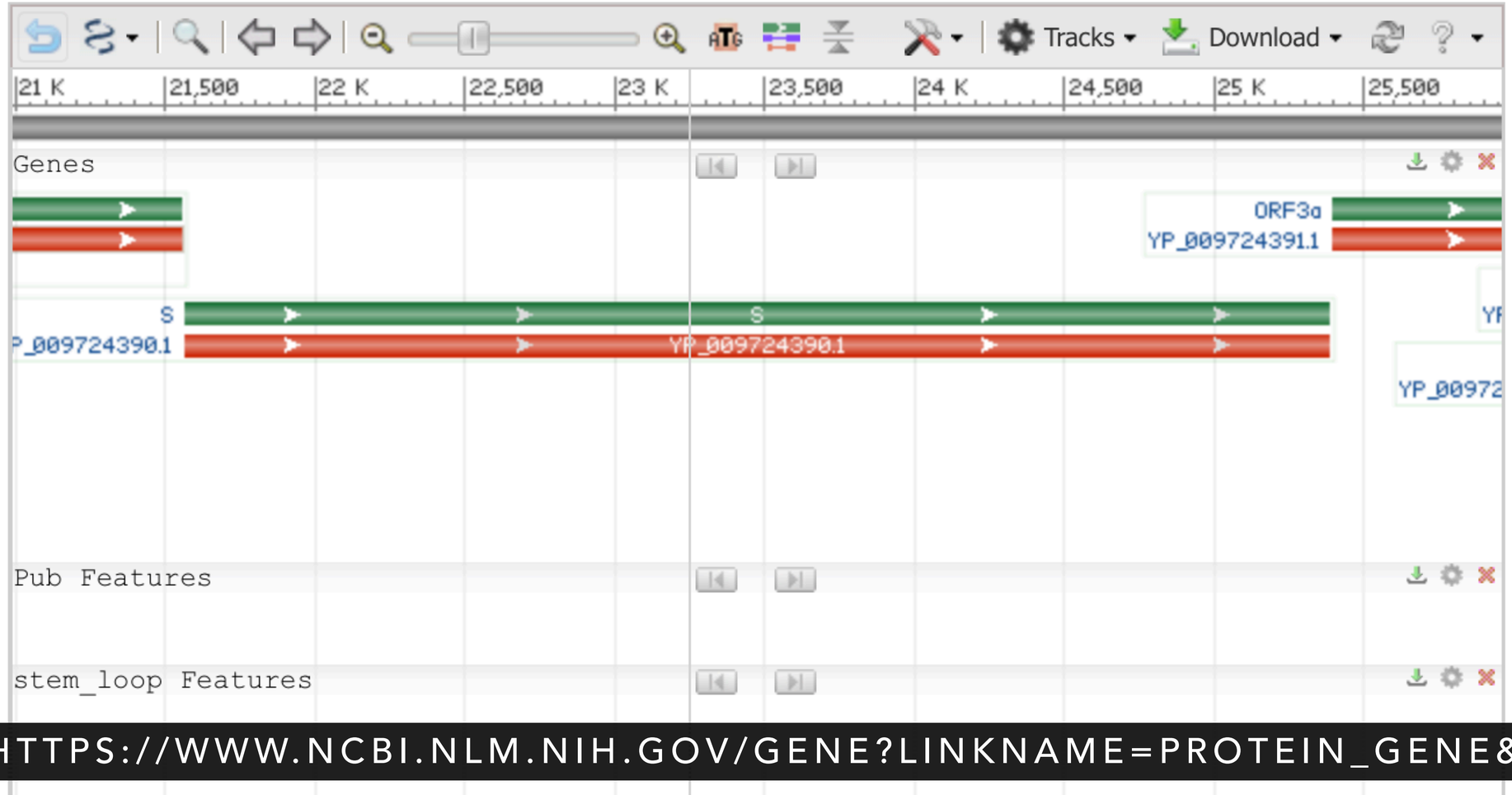
Retrieve, view, and download S
CoV-2 coronavirus genomic an
sequences.

HTTPS://WWW.NCBI.NLM.NIH.GOV/PROTEIN/1791269090

CORONAVIRUS

Genomic Sequence: NC_045512.2

Go to nucleotide: Graphics FASTA GenBank



Taxonomy

General information

About Gene

FAQ

FTP site

Help

My NCBI help

NCBI Handbook

Statistics

Related sites

BLAST

Genome

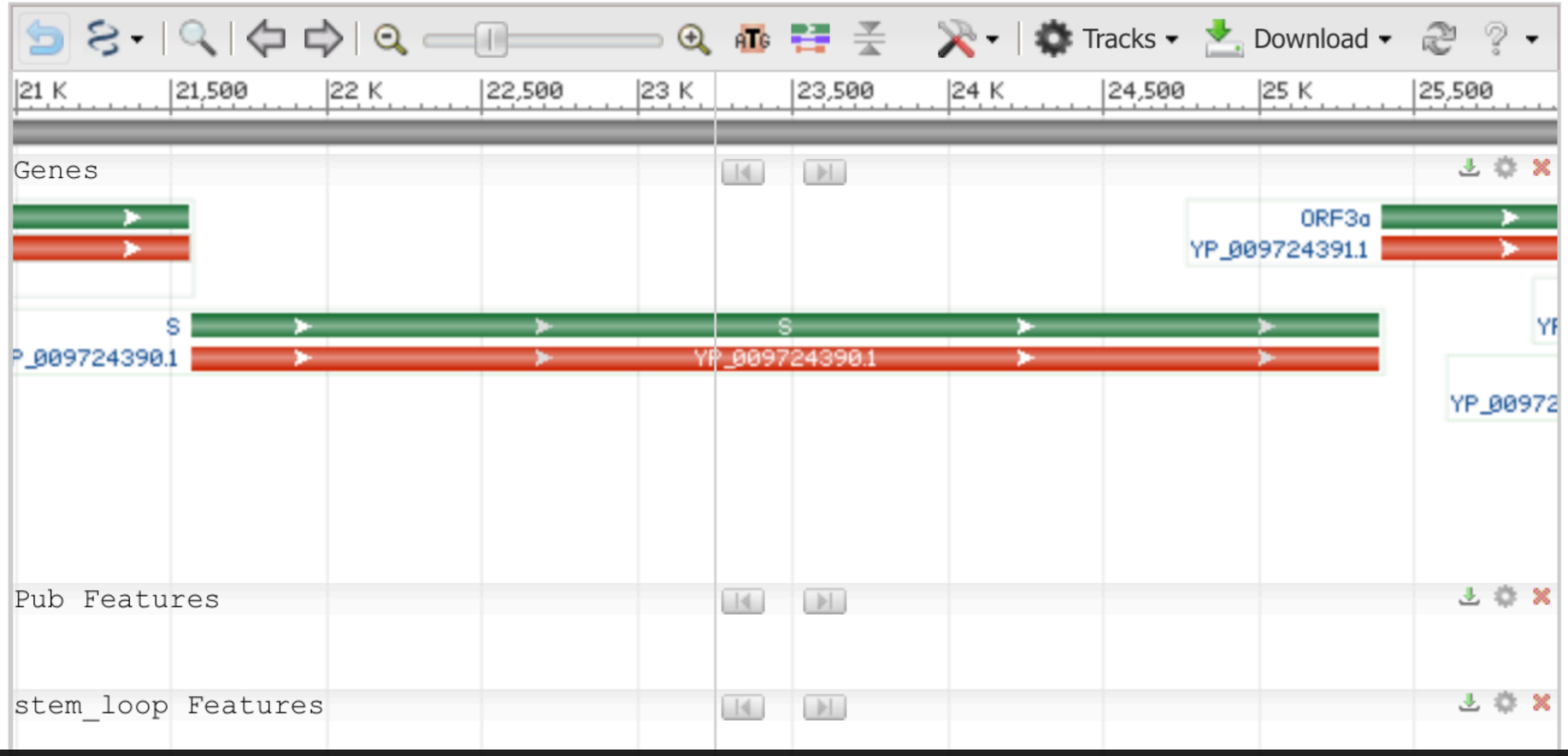
BioProject

HTTPS://WWW.NCBI.NLM.NIH.GOV/GENE?LINKNAME=PROTEIN_GENE&FROM_UID=1791269090

CORONAVIRUS

Genomic Sequence: NC_045512.2

Go to nucleotide: Graphics FASTA GenBank



Taxonomy

General information

About Gene

FAQ

FTP site

Help

My NCBI help

NCBI Handbook

Statistics

Related sites

BLAST

Genome

BioProject

HTTPS://WWW.NCBI.NLM.NIH.GOV/GENE?LINKNAME=PROTEIN_GENE&FROM_UID=1791269090

CORONAVIRUS

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

QHD43416.1

Query subrange

From

To

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

New

Choose Search Set

Database Protein Data Bank proteins(pdb)

HTTPS://WWW.NCBI.NLM.NIH.GOV/GENE?LINKNAME=PROTEIN_GENE&FROM_UID=1791269090

CORONAVIRUS

RID	9H3KDZ07014 *Search expires on 04-18 00:57 am* Download All
Program	BLASTP Citation
Database	pdb See details
Query ID	QHD43416.1
Description	surface glycoprotein [Severe acute respiratory syndrome cor ...
Molecule type	amino acid
Query Length	1273
Other reports	Distance tree of results Multiple alignment MSA viewer

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity	E value	Query Coverage
to	to	to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *77 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Chain A, SARS-CoV-2 spike glycoprotein [Severe acute respiratory syndrome coronavirus 2]	2475	2475	94%	0.0	99.59%	6VSB_A
Chain A, SARS-CoV-2 spike glycoprotein [Severe acute respiratory syndrome coronavirus 2]	2460	2460	94%	0.0	99.50%	6VXX_A
Chain A, SARS-CoV-2 spike glycoprotein [Severe acute respiratory syndrome coronavirus 2]	2458	2458	94%	0.0	99.42%	6VYB_A
Chain A, Spike glycoprotein [Severe acute respiratory syndrome-related coronavirus]	1916	1916	95%	0.0	75.12%	6ACC_A

CORONAVIRUS

Chain A, SARS-CoV-2 spike glycoprotein

PDB: 6VSB_A

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       6VSB_A                  1288 aa            linear   VRL 02-APR-2020
DEFINITION  Chain A, SARS-CoV-2 spike glycoprotein.
ACCESSION   6VSB_A
VERSION     6VSB_A
DBSOURCE    pdb: molecule 6VSB, chain 65, release Apr 1, 2020;
            deposition: Feb 10, 2020;
            class: VIRAL PROTEIN;
            source: Mmdb_id: 184713, Pdb_id 1: 6VSB;
            Exp. method: Electron Microscopy.
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Nidovirales; Cornidovirineae; Coronaviridae;
            Orthocoronavirinae; Betacoronavirus; Sarbecovirus.
REFERENCE   1  (residues 1 to 1288)
  AUTHORS   Wrapp,D., Wang,N., Corbett,K.S., Goldsmith,J.A., Hsieh,C.L.,
            Abiona,O., Graham,B.S. and McLellan,J.S.
  TITLE     Cryo-EM structure of the 2019-nCoV spike in the prefusion
            conformation
  JOURNAL   Science 367 (6483), 1260-1263 (2020)
   PUBMED   32075877
REFERENCE   2  (residues 1 to 1288)
  AUTHORS   Wrapp,D., Wang,N., Corbett,K.S., Goldsmith,J.A., Hsieh,C.,
            Abiona,O., Graham,B.S. and McLellan,J.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-FEB-2020)
COMMENT     Prefusion 2019-nCoV spike glycoprotein with a single
            receptor-binding domain up.
FEATURES             Location/Qualifiers
```

Customize view

Analyze this sequence

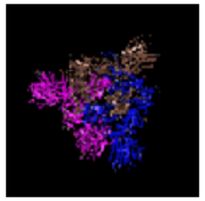
Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure



Prefusion 2019-nCoV spike glycoprotein with a single receptor-binding domain up

PDB: 6VSB

Source: Severe acute respiratory syndrome coronavirus 2

Method: Electron Microscopy

Resolution: 3.46 Å

NCBI Virus

Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Related information

CORONAVIRUS

Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS

Science (2020) **367** p.1260-1263

Abstract

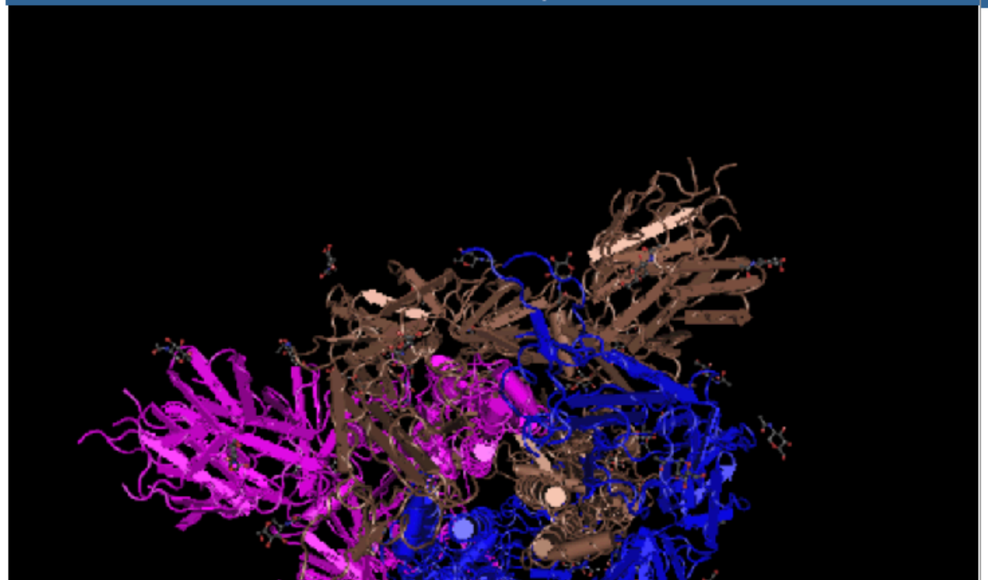
The outbreak of a novel coronavirus (2019-nCoV) represents a pandemic threat that has been declared a public health emergency of international concern. The CoV spike (S) glycoprotein is a key target for vaccines, therapeutic antibodies, and diagnostics. To facilitate medical countermeasure development, we determined a 3.5-angstrom-resolution cryo-electron microscopy structure of... *read more*

PDB Deposition Date: 2020/2/10
Updated in MMDB: 2020/04
Experimental Method: electron micros...
Resolution: 3.46 Å
Source Organism: Severe acut...
Similar Structures: VAST+

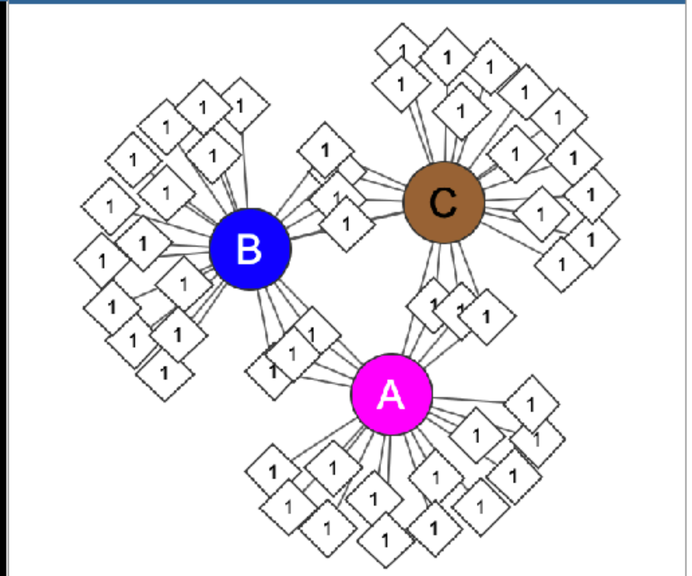
Download sequence data

Biological Unit for 6VSB: trimeric; determined by author

Molecular Graphic ?



Interactions ?



BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
SPRING 2020
SESSION 2

THE UNIVERSITY OF CHICAGO

© T.A. BINKOWSKI, 2020