Nucleotide Databases

Shatakshi Kulkarni

Introduction

❑ Nucleotide databases are specialized repositories that store nucleotide sequences from DNA and RNA.

❑ **Applications:**

- Comparative genomics and evolutionary studies.
- Gene and genome annotation.
- Identifying and characterizing genes and regulatory elements.
- Studying genetic variation and its implications for health and disease.

Primary Nucleotide Databases

Primary Databases

- Primary databases contain raw, unprocessed data directly submitted by researchers.
- These databases typically store original experimental results without additional interpretation or significant processing. E.g., raw nucleotide sequences, raw proteomics data.
- They provide access to unprocessed nucleotide sequences along with associated metadata, such as the source organism and experimental conditions.
- Researchers can submit their nucleotide sequences along with metadata through various submission tools provided by these databases.
- Users can search for sequences based on various criteria such as gene name, organism, or sequence similarity.
- These databases offer advanced search tools and interfaces for data retrieval.
- **Primary nucleotide databases:**

A. GenBank

B. ENA

C. ARSA

GenBank

- GenBank (R) is a comprehensive database that contains publicly available nucleotide sequences for more than 240 000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects.
- Submissions made using web based BankIt or Sequin programs.
- GenBank is accessible through NCBI's retrieval system, Entrez.
- Key divisions in GenBank:

a. WGS: Major portion of database consists of whole genome sequencing i.e., complete genetic material of an organism

b. EST/Expressed sequence tags: About 16% of the sequences in GenBank are of human origin and 13% of all sequences are human ESTs. ESTs are short sub-sequences of cDNA sequences which helps to understand the gene expression in specific tissues.

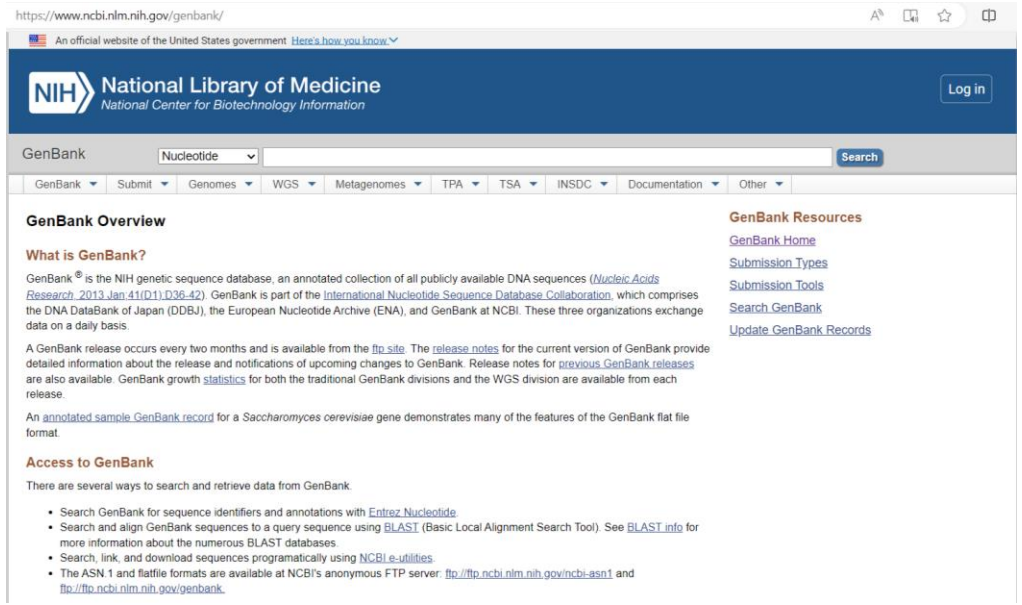
c. High Throughput Genomic (HTG): contains sequences generated from large-scale genomic projects through sequencing technologies. HTG supports genome assembly, mapping and annotation of these sequences.

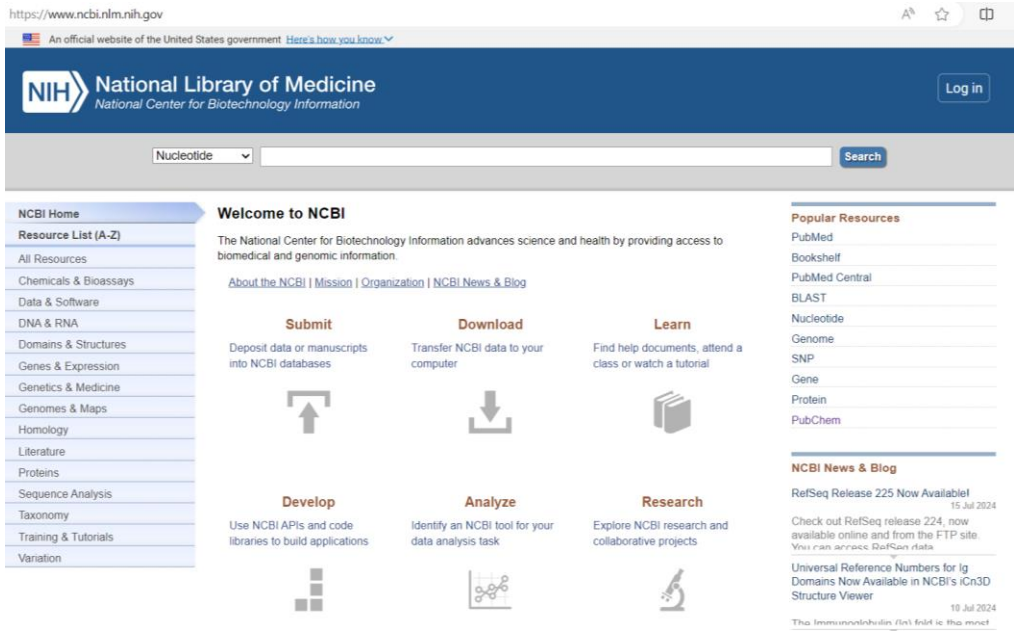
d. ENV: The ENV division contains sequences obtained from environmental samples, which can include sequences from microbial communities in various environments (e.g., soil, water, human gut). These sequences are used in metagenomics data.

e. PRI/Primates, ROD/Rodents, BCT/bacteria, MAM/Other mammals, etc.

GenBank Homepage

- GenBank Overview (nih.gov) or National Center for Biotechnology Information (nih.gov) and choose nucleotide



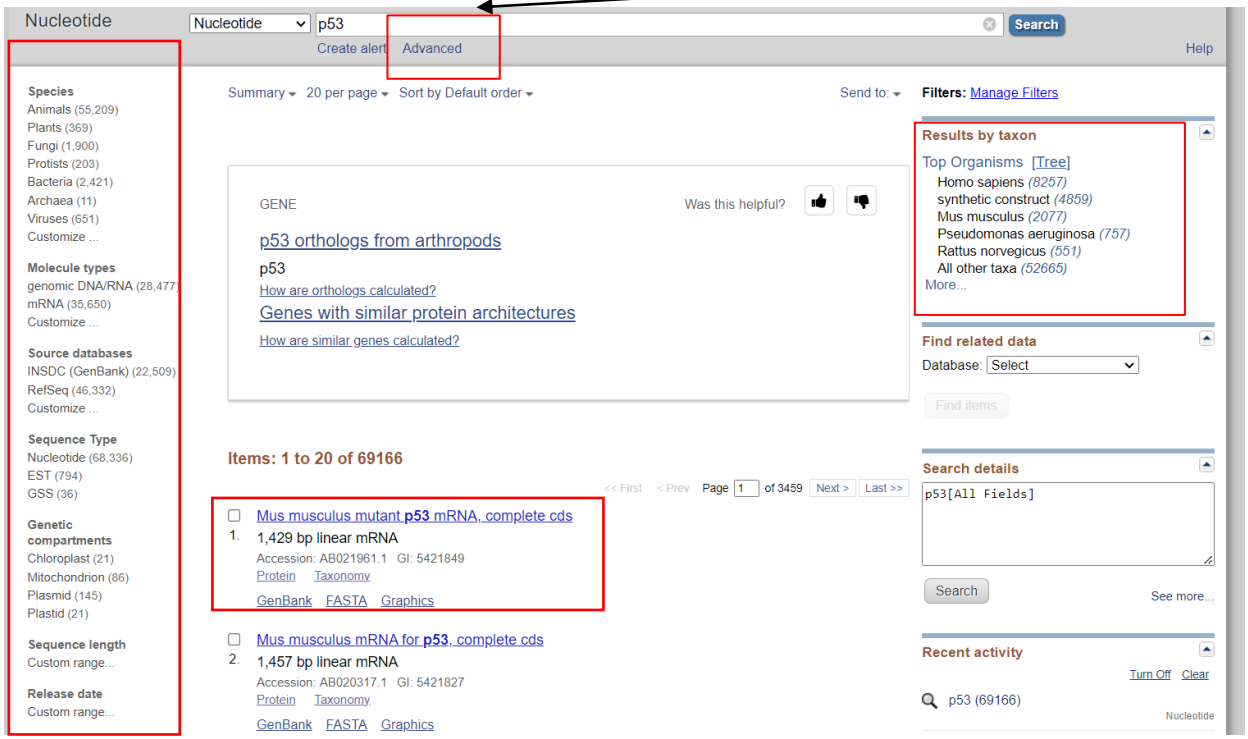


GenBank Query Search

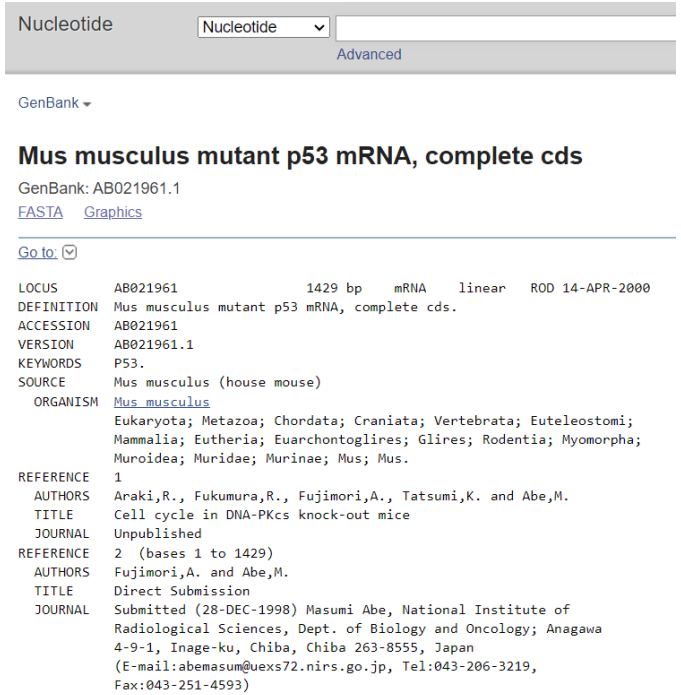
- Organism, name, no. of nucleotides and accession number (unique identifier of a sequence)

Advanced Filtered Search

Filter



GenBank Flat file

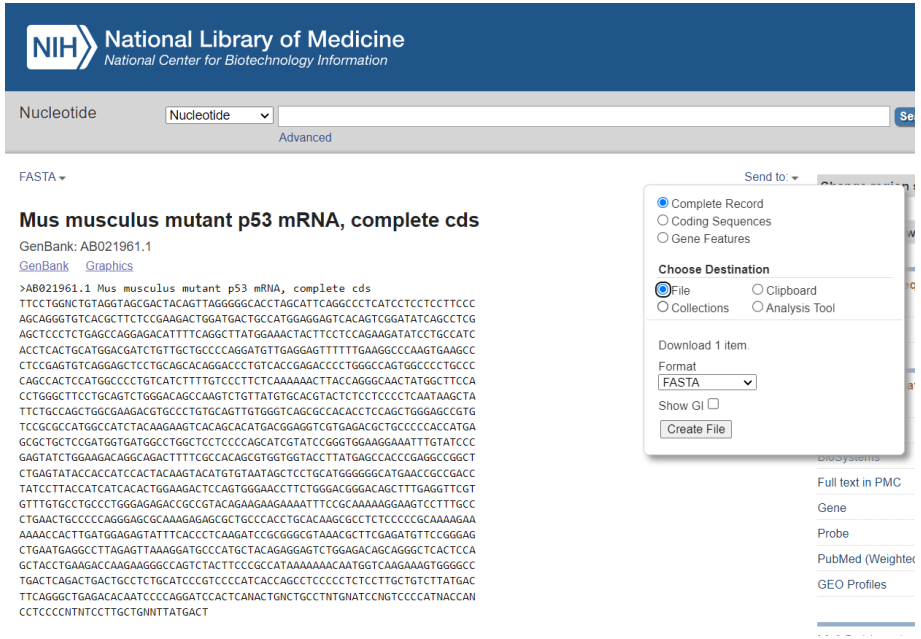



```
FEATURES             Location/Qualifiers
     source          1..829
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     gene            <1..>695
                     /gene="HBA1"
     mRNA            join(<1..95,213..417,567..>695)
                     /gene="HBA1"
                     /product="hemoglobin alpha 1"
     CDS             join(1..95,213..417,567..695)
                     /gene="HBA1"
                     /codon_start=1
                     /product="hemoglobin alpha 1"
                     /protein_id="AFI57164.1"
                     /translation="MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYF
                     PHFDLSHGSAQVKGHGKKVADALTNAVAHVHDMPNALSALSDLHAHKLRVDPVNFKLL
                     SHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR"
     variation       340
                     /gene="HBA1"
                     /note="Asp75His; results in histidine to aspartic acid
                     substitution"
                     /replace="g"
```

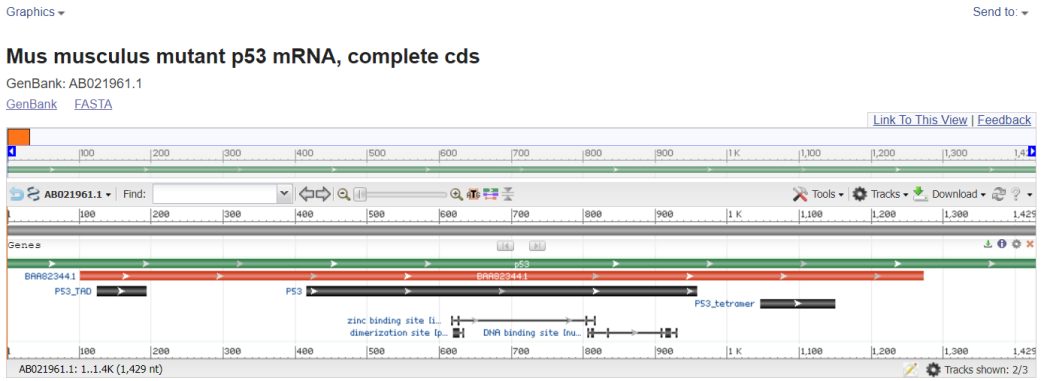
```
ORIGIN
        1 atggtgctgt ctcctgccga caagaccaac gtcaaggccg cctggggtaa ggtcggcgcg
       61 cacgctggcg agtatggtgc ggaggccctg gagaggtgag gctccctccc ctgctccgac
      121 ccgggctcct cgcccgcccg gacccacagg ccaccctcaa ccgtcctggc cccggaccca
      181 aaccccaccc ctcactctgc ttctccccgc aggatgttcc tgtccttccc caccaccaag
      241 acctacttcc cgcacttcga cctgagccac ggctctgccc aggttaaggg ccacggcaag
      301 aaggtggccg acgcgctgac caacgccgtg gcgcacgtgc acgacatgcc caacgcgctg
      361 tccgccctga gcgacctgca cgcgcacaag cttcgggtgg acccggtcaa cttcaaggtg
      421 agcggcgggc cgggagcgat ctgggtcgag gggcgagatg gcgccttcct cgcagggcag
      481 aggatcacgc gggttgcggg aggtgtagcg caggcggcgg ctgcgggcct gggccctcgg
      541 ccccactgac cctcttctct gcacagctcc taagccactg cctgctggtg accctggccg
      601 cccacctccc cgccgagttc acccctgcgg tgcacgcctc cctggacaag ttcctggctt
      661 ctgtgagcac cgtgctgacc tccaaatacc gttaagctgg agcctcggtg gccatgcttc
      721 ttgccccttg ggcctccccc cagcccctcc tccccttcct gcacccgtac ccccgtggtc
      781 tttgaataaa gtctgagtgg gcggcagcct gtgtgtgcct gagtttttt
```

FASTA

- FASTA format: Nucleotides and Amino acids presented using single letter codes



FASTA



Graphics View

GenBank Flat file

- **Locus:** Accession number, no. of nucleotides/base pairs, type of molecule, topology, division and date of submission
- **Definition:** Source/organism, name of gene
- **Features:** genes and gene products along with regions of biological significance with their locations. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features. Introns, exons can also be included.
  - Source

    /organism: Will specify the organism from which sequence is derived.

    /mol_type: indicates molecular type of sequence

    /db_xref: Cross-reference to taxonomy database
  - Gene:

    /gene specifies gene symbol

GenBank Flat file

❑ **Features (Con):**

- mRNA: Describes the locations of exons that are joined together to form the mature mRNA transcript.

  /gene provides gene associated with mRNA

  /product Specifies the product of this mRNA

- CDS: Coding sequence responsible for coding for a protein. /translation provides the protein that gene sequence is coding for.

- Variation:

  /note Describes the nature of the variation. E.g., amino acid substitution

  /replace Specifies the nucleotide change responsible for the variation.

GenBank Flat file

❑ **Features (Con):**

Locations in Feature section:

- 1...1429

Complete feature (Whole sequence)

- <1..>695

The < symbol indicates partial on the 5' end.

The > symbol indicates partial on the 3' end.

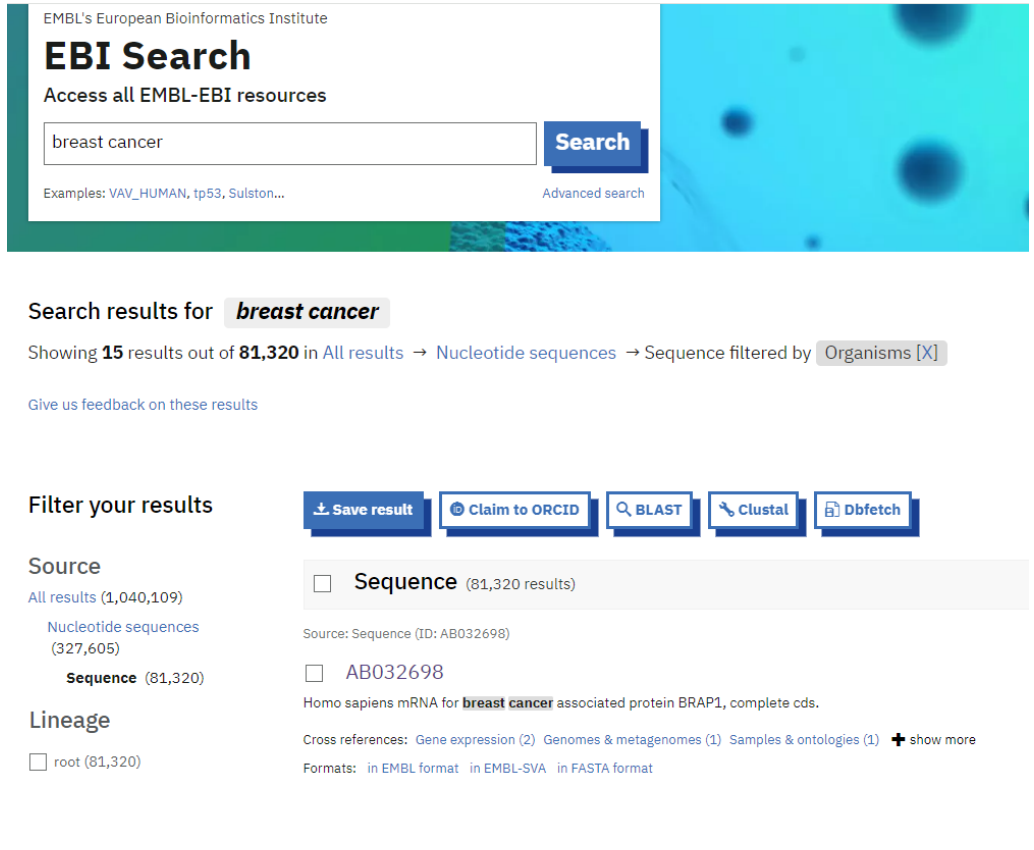
❑ **Origin:** nucleotide sequence of the gene

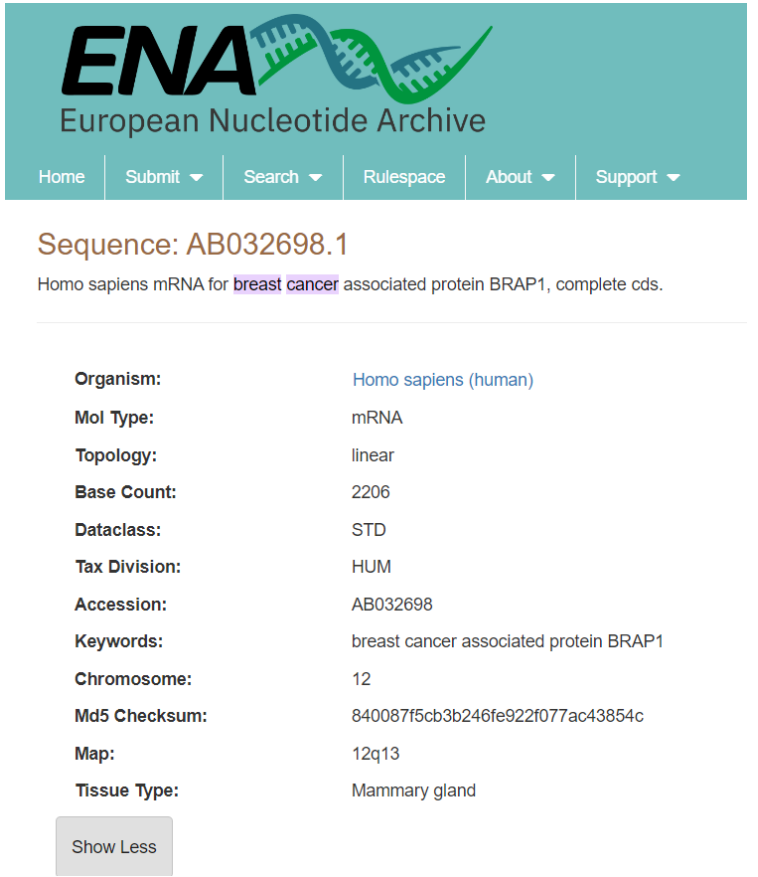
ENA/European Nucleotide Archive

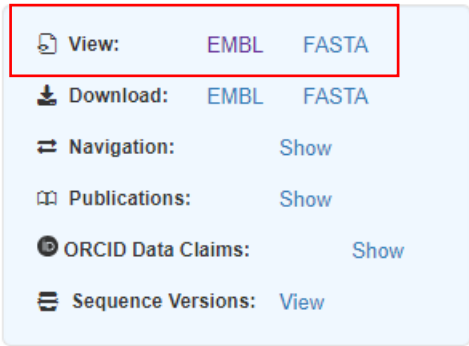
- https://www.ebi.ac.uk/ and choose nucleotide sequences (Filter Sequences)
- A comprehensive repository that provides access to primary nucleotide sequences
- Accepts data in various formats, including FASTQ for raw reads, FASTA for sequences, and XML for metadata
- ENA data is cross-referenced with other major biological databases, including GenBank, DDBJ, UniProt, and the Protein Data Bank (PDB)
- Data can be used to further perform analyses such as alignment, variant calling, or functional annotation
- Widely used by researchers for accessing reference genomes, comparative genomics, and evolutionary studies
- Flat file contains a ladder of information and total nucleotide count

ENA

Can Filter the search according to ENA (Sequences) and organisms







ENA Flat File

Accession No., Topology, Genomic type, Organism, Total nucleotide bases

Date created & updated

Description

Taxonomy (Organism source)

Reference No., Reference Author Reference Title Reference Literature

Cross-references

```
ID   AB032698; SV 1; linear; mRNA; STD; HUM; 2206 BP.
XX
AC   AB032698;
XX
DT   10-DEC-1999 (Rel. 62, Created)
DT   07-OCT-2008 (Rel. 97, Last updated, Version 3)
XX
DE   Homo sapiens mRNA for breast cancer associated protein BRAP1, complete cds.
XX
KW   breast cancer associated protein BRAP1.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-2206
RA   Miki Y., Saito H.;
RT   ;
RL   Submitted (22-SEP-1999) to the INSDC.
RL   Yoshio Miki, The Cancer Institute, Japanese Foundation for Cancer Research,
RL   Dept. of Molecular Diagnosis; 1-37-1, Kami-ikebukuro, Toshimaku, Tokyo
RL   170-8455, Japan (E-mail:yosmiki@ims.u-tokyo.ac.jp, Tel:81-3-5394-4035,
RL   Fax:81-3-5394-4035)
XX
RN   [2]
RP   1-2206
RA   Miki Y., Saito H.;
RT   "Characterization of a novel breast cancer associated protein BRAP1";
RL   Unpublished.
XX
DR   MD5; 840087f5cb3b246fe922f077ac43854c.
DR   Ensembl-Gn; ENSG00000110934; homo_sapiens.
DR   Ensembl-Tr; ENST00000544402; homo_sapiens.
XX
```

ENA Flat File

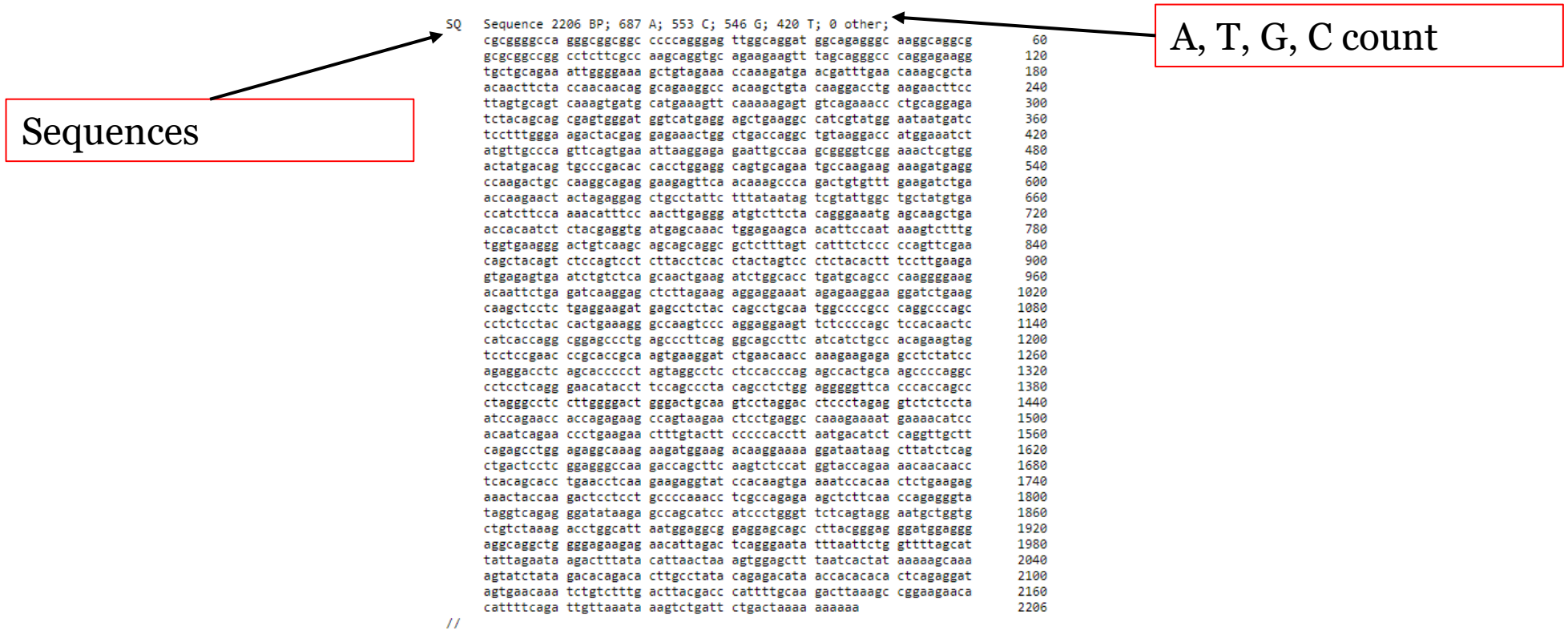
```
FH   Key             Location/Qualifiers
FH
FT   source          1..2206
FT                   /organism="Homo sapiens"
FT                   /chromosome="12"
FT                   /map="12q13"
FT                   /mol_type="mRNA"
FT                   /tissue_type="Mammary gland"
FT                   /db_xref="taxon:9606"
FT   exon            1..119
FT                   /number=1
FT   CDS             39..1736
FT                   /codon_start=1
FT                   /transl_table=1
FT                   /gene="BRAP1"
FT                   /product="breast cancer associated protein BRAP1"
FT                   /note="a novel protein related to BIN1, amphiphysin and RVS
FT                   (BAR) family"
FT                   /db_xref="GOA:Q9UBW5"
FT                   /db_xref="H-InvDB:HIT000059063.12"
FT                   /db_xref="HGNC:HGNC:1053"
FT                   /db_xref="InterPro:IPR003005"
FT                   /db_xref="InterPro:IPR004148"
FT                   /db_xref="InterPro:IPR027267"
FT                   /db_xref="PDB:4AVM"
FT                   /db_xref="PDB:4I1Q"
FT                   /db_xref="UniProtKB/Swiss-Prot:Q9UBW5"
FT                   /protein_id="BAA88108.1"
FT                   /translation="MAEGKAGGAAGLFAKQVQKKFSRAQEKVLQKLGKAVETKDERFEQ
FT                   SANNFYQQQAEGHKLYKDLKNFLSAVKVMHESSKRVSETLQEIYSSEWDGHEELKAIVW
FT                   NNDLLWEDYEEKLADQAVRTMEIYVAQFSEIKERIAKRGRKLVDYDSARHHLEAVQNAK
FT                   KKDEAKTAKAEEEFNKAQTVFEDLNQELLEELPILYNSRIGCYVTIFQNISNLRDVFYR
FT                   EMSKLNHNLYEVMSKLEKQHSNKVFVVKGLSSSSRRSLVISPPVRTATVSSPLTSPTSP
FT                   STLSLKSESESVSATEDLAPDAAQGEDNSEIKELLEEEEIEKEGSEASSSEEDEPLPAC
FT                   NGPAQAQPSPTTERAKSQEEVLPSSTTPSPGGALSPSGQPSSSATEVVLRTRTASEGSE
FT                   QPKKRASIQRTSAPPSRPPPPRATASPRPSSGNIPSSPTASGGGSPTSPRASLGTGTAS
FT                   PRTSLEVSPNPEPPEKPVRTPEAKENENIHNQNPEELCTSPTLMTSQVASEPGEAKKME
FT                   DKEKDNKLISADSSEGQDQLQVSMVPENNNLTAPEPQEEVSTSENPQL"
```

Cross References

AA sequence (Protein) coded by CDS

```
FT   exon            120..200
FT                   /number=2
FT   exon            201..255
FT                   /number=3
FT   exon            256..350
FT                   /number=4
FT   exon            351..446
FT                   /number=5
FT   exon            447..554
FT                   /number=6
FT   exon            555..640
FT                   /number=7
FT   exon            641..716
FT                   /number=8
FT   exon            717..799
FT                   /number=9
FT   exon            800..1553
FT                   /number=10
FT   exon            1554..1634
FT                   /number=11
FT   exon            1635..1706
FT                   /number=12
FT   exon            1707..2206
FT                   /number=13
XX
```

ENA Flat File




```
SQ   Sequence 2206 BP; 687 A; 553 C; 546 G; 420 T; 0 other;
     cgcggggcca gggcggcggc ccccagggag ttggcaggat ggcagagggc aaggcaggcg        60
     gcgcggccgg cctctttcgcc aagcaggtgc agaagaagtt tagcagggcc caggagaagg       120
     tgctgcagaa attggggaaa gctgtagaaa ccaaagatga acgatttgaa caaagcgcta       180
     acaacttcta ccaacaacag gcagaaggcc acaagctgta caaggacctg aagaacttcc       240
     ttagtgcagt caaagtgatg catgaaagtt caaaaagagt gtcagaaacc ctgcaggaga       300
     tctacagcag cgagtgggat ggtcatgagg agctgaaggc catcgtatgg aataatgatc       360
     tcctttggga agactacgag gagaaactgg ctgaccaggc tgtaaggacc atggaaatct       420
     atgttgccca gttcagtgaa attaaggaga gaattgccaa gcggggtcgg aaactcgtgg       480
     actatgacag tgcccgacac cacctggagg cagtgcagaa tgccaagaag aaagatgagg       540
     ccaagactgc caaggcagag gaagagttca acaaagccca gactgtgttt gaagatctga       600
     accaagaact actagaggag ctgcctattc tttataatag tcgtattggc tgctatgtga       660
     ccatcttcca aaacatttcc aacttgaggg atgtcttcta cagggaaatg agcaagctga       720
     accacaatct ctacgaggtg atgagcaaac tggagaagca acattccaat aaagtctttg       780
     tggtgaaggg actgtcaagc agcagcaggc gctctttagt catttctccc ccagttcgaa       840
     cagctacagt ctccagtcct cttacctcac ctactagtcc ctctacactt tccttgaaga       900
     gtgagagtga atctgtctca gcaactgaag atctggcacc tgatgcagcc caaggggaag       960
     acaattctga gatcaaggag ctcttagaag aggaggaaat agagaaggaa ggatctgaag      1020
     caagctcctc tgaggaagat gagcctctac cagcctgcaa tggccccgcc caggcccagc      1080
     cctctcctac cactgaaagg gccaagtccc aggaggaagt tctccccagc tccacaactc      1140
     catcaccagg cggagccctg agcccttcag ggcagccttc atcatctgcc acagaagtag      1200
     tcctccgaac ccgcaccgca agtgaaggat ctgaacaacc aaagaagaga gcctctatcc      1260
     agaggacctc agcacccccc agtaggcctc ctccacccag agccactgca agccccaggc      1320
     cctcctcagg gaacatacct tccagcccta cagcctctgg agggggttca cccaccagcc      1380
     ctagggcctc cttggggact gggactgcaa gtcctaggac ctccctagag gtctctccta      1440
     atccagaacc accagagaag ccagtaagaa ctcctgaggc caaagaaaat gaaaacatcc      1500
     acaatcagaa ccctgaagaa ctttgtactt cccccacctt aatgacatct caggttgctt      1560
     cagagcctgg agaggcaaag aagatggaag acaaggaaaa ggataataag cttatctcag      1620
     ctgactcctc ggagggccaa gaccagcttc aagtctccat ggtaccagaa aacaacaacc      1680
     tcacagcacc tgaacctcaa gaagaggtat ccacaagtga aaatccacaa ctctgaagag      1740
     aaactaccaa gactcctcct gccccaaacc tcgccagaga agctcttcaa ccagagggta      1800
     taggtcagag ggatataaga gccagcatcc atccctgggt tctcagtagg aatgctggtg      1860
     ctgtctaaag acctggcatt aatggaggcg gaggagcagc cttacgggag ggatggaggg      1920
     aggcaggctg gggagaagag aacattagac tcagggaata tttaattctg gttttagcat      1980
     tattagaata agactttata cattaactaa agtggagctt taatcactat aaaaagcaaa      2040
     agtatctata gacacagaca cttgcctata cagagacata accacacaca ctcagaggat      2100
     agtgaacaaa tctgtctttg acttacgacc cattttgcaa gacttaaagc cggaagaaca      2160
     cattttcaga ttgttaaata aagtctgatt ctgactaaaa aaaaaa                     2206
//
```

ARSA

- ARSA | DDBJ | Quick Search (nig.ac.jp)
- Primary nucleotide sequence database stored in DDBJ databases
- ARSA supports the use of Boolean operators (AND, OR, NOT) to refine search queries.
- FASTA, Flatfile can be downloaded
- DDBJ collaborates with GenBank and ENA to ensure that nucleotide sequence data is shared and updated across all three databases.
- Sequence records in DDBJ are cross-referenced with related entries in other biological databases, enhancing data interoperability.
- Flat file provides total nucleotide count, detailed information about variation

ARSA

```
LOCUS       FW351573                62 bp    DNA     linear   PAT 24-AUG-2012
DEFINITION  WO 2010058819-A/10: Peptide inhibiting interaction between human
            cancer protein MDM2 and human cancer suppression protein p53, and
            use thereof.
ACCESSION   FW351573
VERSION     FW351573.1
KEYWORDS    WO 2010058819-A/10.
SOURCE      synthetic construct
  ORGANISM  synthetic construct
            other sequences; artificial sequences.
REFERENCE   1  (bases 1 to 62)
  AUTHORS   Yanagawa,H. and Shiheido,H.
  TITLE     Peptide inhibiting interaction between human cancer protein MDM2
            and human cancer suppression protein p53, and use thereof
  JOURNAL   Patent: WO 2010058819-A 10 27-MAY-2010;
            Keio University
COMMENT     OS   Artificial sequence
            PN   WO 2010058819-A/10
            PD   27-MAY-2010
            PF   19-NOV-2009 WO 2009JP069644
            PR   19-NOV-2008 JP 200 8-296172
            PA   Keio University
            PI   hiroshi yanagawa,hirokazu shiheido
            PT   "Peptide inhibiting interaction between human cancer protein
            PT   MDM2 and human
            PT   cancer suppression protein p53, and use thereof"
            PS   N15
            CC   primer priSP6OGf
            FH   Key             Location/Qualifiers
FEATURES             Location/Qualifiers
     source          1..62
                     /mol_type="unassigned DNA"
                     /db_xref="taxon:32630"
                     /organism="synthetic construct"
BASE COUNT           27 a           12 c           13 g           10 t
ORIGIN
        1 atttaggtga cactatagaa caacaacaac aacaaacaac aacaaaatgg gtggcggcgg
       61 tt
//
```