Biological Databases

Dr. Upeksha Ganegoda

Department of Computational Mathematics

Outline

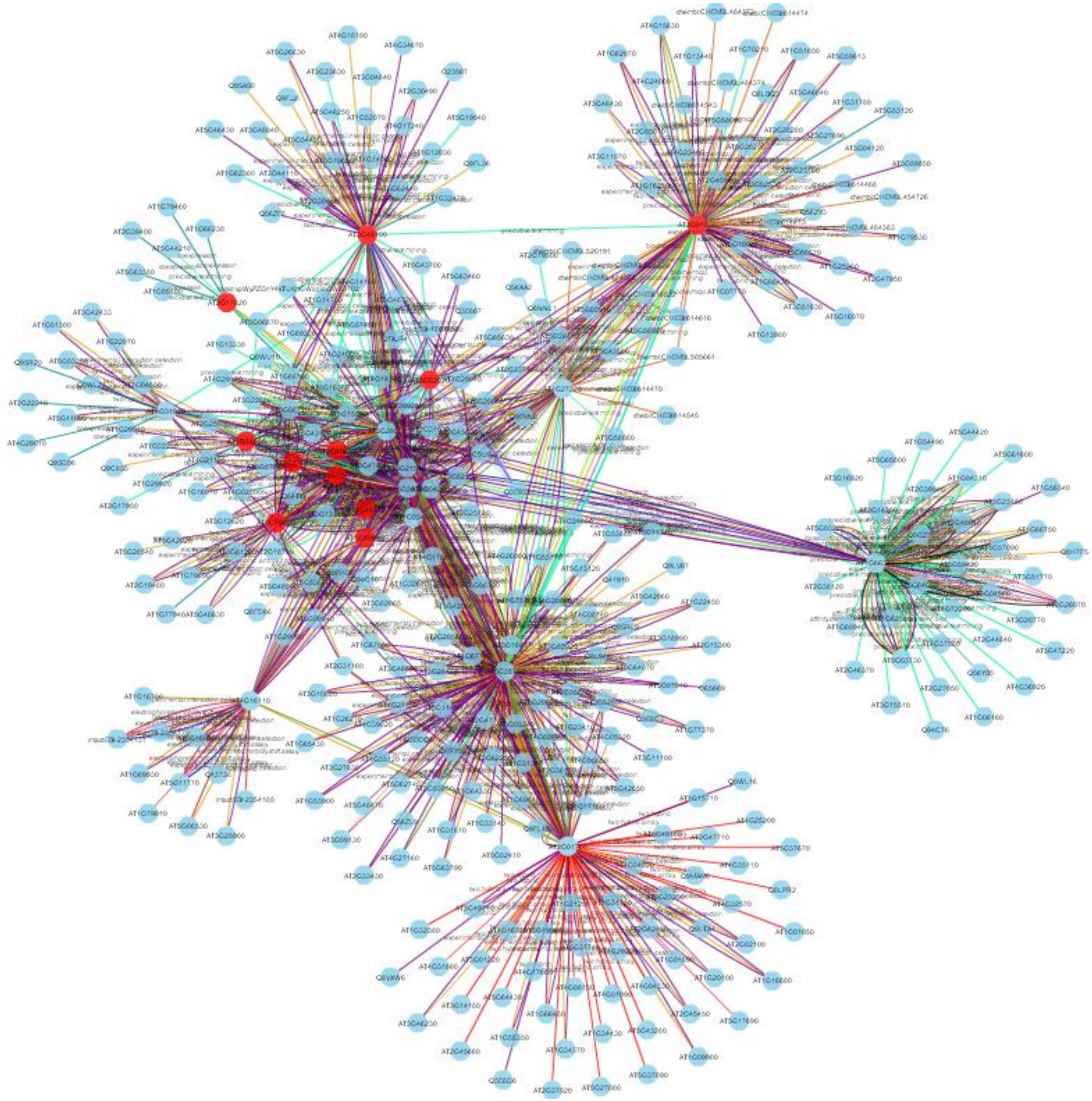
- Different types of biological networks
- Database Structures
- Biological database types based on content

Biological Networks

- Protein-protein interaction network
- Metabolic network
- Gene regulatory network
- RNA network

Protein-protein interaction network

- A protein can interact with another protein, in order to build a protein complex or to activate it. By using a protein-protein interaction network it shows how and which proteins are interact each other.
- Node represent a protein, Arc represent the interaction between two protein.
- Can use different types of graph algorithms to identify:
  - Protein complexes
  - Protein functions
  - Protein Hubs
  - etc



Protein Hub?

Protein complexes?

Metabolic network

- Metabolic networks give an in-depth insight of the molecular mechanisms of a particular organism. It will correlate the genome with molecular physiology and provide the most comprehensive of all biological networks.

  Ex: Databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Biochemical Genetic and Genomics knowledgebase (BIGG) contain the metabolic network of a wide range of species.

Gene-regulatory network

- It is a common type of regulatory network
- gene regulatory network consist of DNA segments in a cell which interact with each other indirectly (by using their RNA and protein expression products) and with other materials in the cell to manage the gene expression levels of mRNA and proteins.

RNA networks

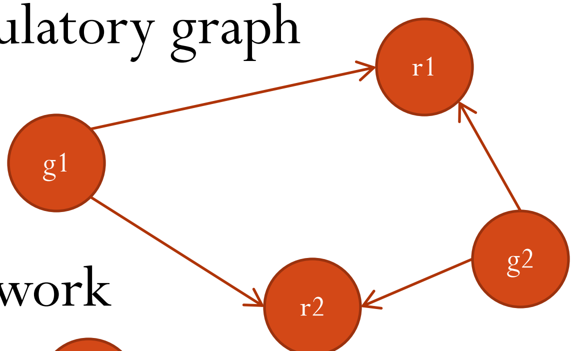
- RNA networks show the interaction between RNA-RNA or RNA-DNA interactions. By understanding the microRNA’s role in disease, the researchers able to construct microRNA-gene networks by using predicted microRNA targets available in public databases such as Target Scan, PicTar, microRNA, miRBase and miRDB.

Representation as a network

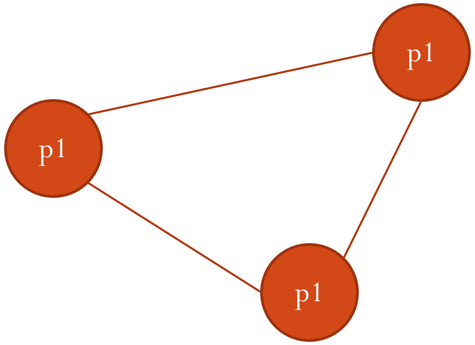
Network G = (V, E, w), where V represents the set of proteins, E is the set of interactions and w denotes the weight of each interaction

Network can construct as

- Directional graph Ex: gene-regulatory graph



- Bidirectional graph Ex: PPI network



Main functions of biological databases

- **Make biological data available to scientists.**
  As much as possible of a particular type of information should be available in one single place (book, site, database). Published data may be difficult to find or access, and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences).
- **To make biological data available in computer-readable form.**
  Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

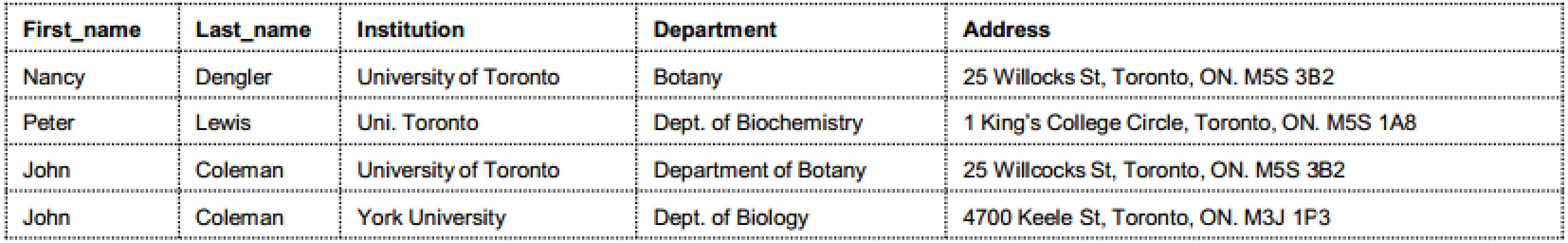
What is a database?

- How can data be stored…

Flat-file format, with fields separated by some delimiter

Nancy|Dengler|Botany|University of Toronto|25 Willocks St, Toronto, ON. M5S 3B2
Peter|Lewis|Dept. of Biochemistry|Uni. Toronto|1 King's College Circle, Toronto, ON. M5S 1A8
John|Coleman|Department of Botany|University of Toronto|25 Willcocks St, Toronto, ON. M5S 3B2
John|Coleman|Dept. of Biology|York University|4700 Keele St, Toronto, ON. M3J 1P3

These data could also be stored in a spreadsheet



First_name	Last_name	Institution	Department	Address
Nancy	Dengler	University of Toronto	Botany	25 Willocks St, Toronto, ON. M5S 3B2
Peter	Lewis	Uni. Toronto	Dept. of Biochemistry	1 King's College Circle, Toronto, ON. M5S 1A8
John	Coleman	University of Toronto	Department of Botany	25 Willcocks St, Toronto, ON. M5S 3B2
John	Coleman	York University	Dept. of Biology	4700 Keele St, Toronto, ON. M3J 1P3

What are the problems with this sort of database?
Relational Databases offer a solution…

Database structures

- Flat files
- Relational
- Object oriented

Relational database

Nancy | Dengler | Botany | University of Toronto | 25 Willocks St, Toronto, ON. M5S 3B2
Peter | Lewis | Dept. of Biochemistry | Uni. Toronto | 1 King’s College Circle, Toronto, ON. M5S 1A8
John | Coleman | Department of Botany | University of Toronto | 25 Willcocks St, Toronto, ON. M5S 3B2
John | Coleman | Dept. of Biology | York University | 4700 Keele St, Toronto, ON. M3J 1P3

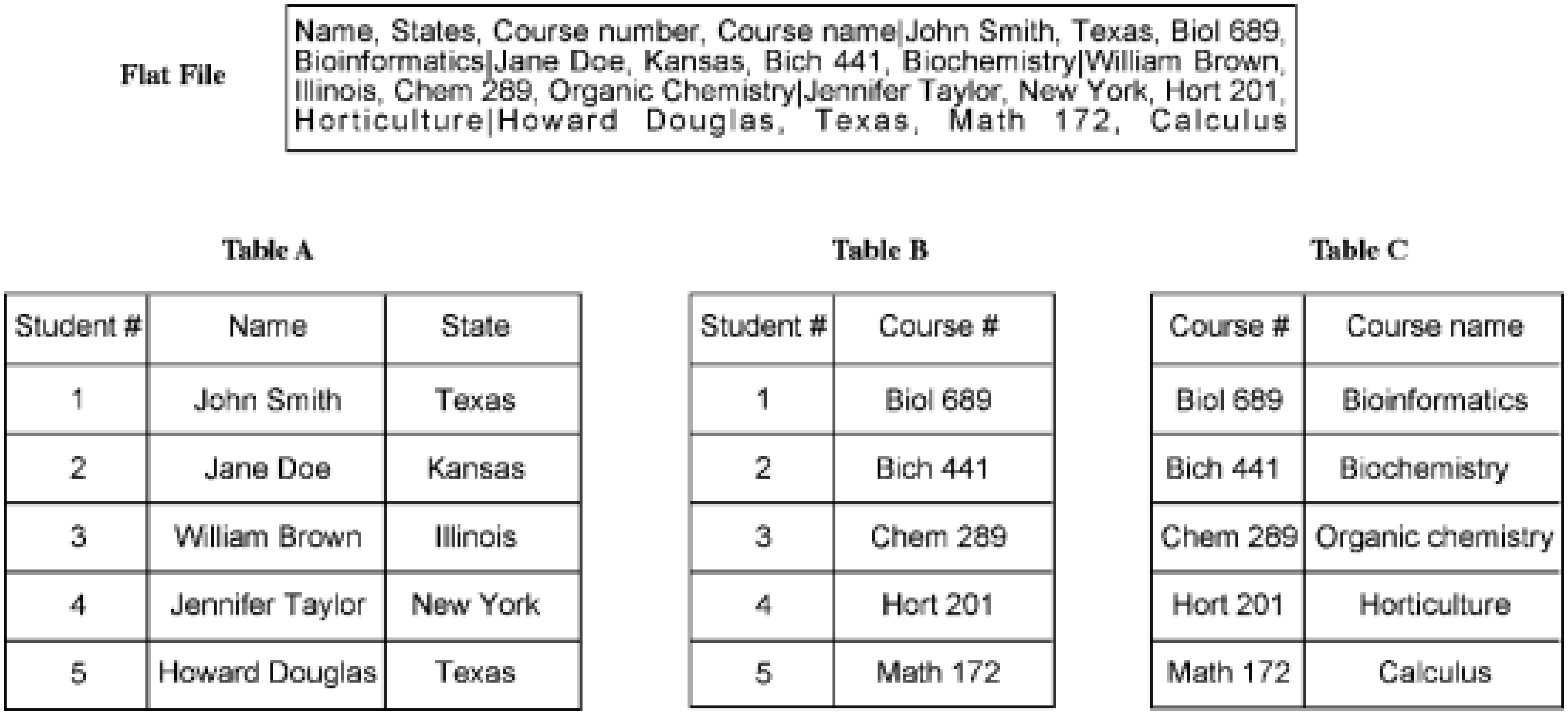
A relational database consists of a relations (tables) containing attributes (fields or columns). Each row in a table is known as a tuple or a record. Information should be ‘normalized’ so that it is non-redundant this means that every row should be unique, although this ideal is not always observed.

Table 'Professors'

Professor_id	First_name	Last_name	Contact_id
1	Nancy	Dengler	1
2	Peter	Lewis	2
3	John	Coleman	1
4	John	Coleman	3

Table 'Contacts'

Contact_id	Institution	Department	Address
1	University of Toronto	Dept. of Botany	25 Willocks St, Toronto, ON. M5S 3B2
2	Uni. Toronto	Dept. of Biochemisty	1 King’s College Circle, Toronto, ON. M5S 1A8
3	York University	Dept. of Biology	4700 Keele St, Toronto, ON. M3J 1P



Flat File

Name, States, Course number, Course name|John Smith, Texas, Biol 689, Bioinformatics|Jane Doe, Kansas, Bich 441, Biochemistry|William Brown, Illinois, Chem 289, Organic Chemistry|Jennifer Taylor, New York, Hort 201, Horticulture|Howard Douglas, Texas, Math 172, Calculus

Table A

Student #	Name	State
1	John Smith	Texas
2	Jane Doe	Kansas
3	William Brown	Illinois
4	Jennifer Taylor	New York
5	Howard Douglas	Texas

Table B

Student #	Course #
1	Biol 689
2	Bich 441
3	Chem 289
4	Hort 201
5	Math 172

Table C

Course #	Course name
Biol 689	Bioinformatics
Bich 441	Biochemistry
Chem 289	Organic chemistry
Hort 201	Horticulture
Math 172	Calculus

Different Database Types

- Primary databases

  Contain original biological data. Ex. Raw nucleic acid sequence data from GeneBank, EMBL database, DNA Data Bank.

- Secondary databases

  Contain computationally processed or manually curated information based on original information from primary database. Ex. SWISS-PROT, TrEMBL (contain translated nucleic acid sequences), PIR (contain annotated protein sequences).

- Specialized databases
  - This will cater to a particular research interest. Ex. Flybase, WormBase, AceDB, and TAIR

Pitfalls of biological databases

- Overreliance of sequence information without understanding the reliability of the information.
- High level of redundancy
- Annotations of genes can occasionally be false or incomplete.

Accession codes, identifiers

- Many of the biological databases (GenBank, UNIPROT etc.) have two (or more!) different ways of identifying a given entry:
  - Identifier
  - Accession code (or number)

- **Identifier**

  An identifier ("locus" in GenBank, "entry name" in UNIPROT) is a string of letters and digits that understandable in some meaningful way by a human.

Identifiers are not as stable as accession numbers, mainly because they are modified by the curators if the presumed function of the protein is found to be something else.

UNIPROT: B5YME7
GenBank: XM_002295694

An identifier can change. For example, the database curators may decide that the identifier for an entry no longer is appropriate. This can happen very rarely.

- **Accession code (number)**

  An accession code (or number) is a number (with a few characters in front) that uniquely identifies an entry. It is often assigned arbitrarily. For example, the accession code for **B5YME7_THAPS** in UNIPROT is **B5YME7**.

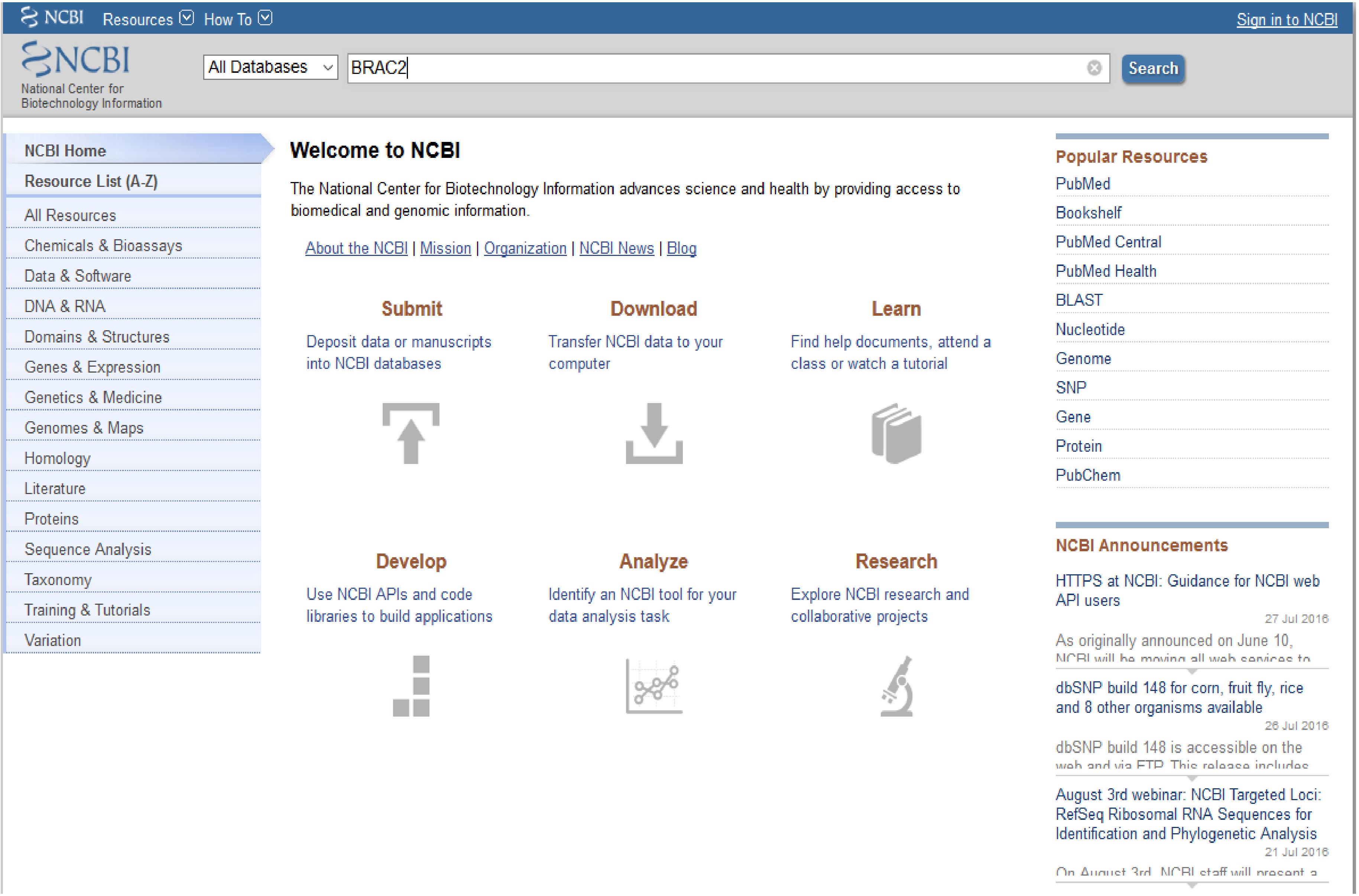
  In the case of GenBank, the accession code for the human BRAC2 gene sequence is XM_002295694.

Versions and Gene Indices

In 1992, NCBI began assigning a unique number for each sequence submitted – the GenInfo Identifier (GI) number. The same accession number may be associated with a different GI if a newer or corrected sequence is submitted.

Records typically contain the Accession.Version identifier, such as XM_002295694.1, in the VERSION line of the record. This identifier is mapped to its unique corresponding GI number, which is the "primary key" of GenBank.

To specify a sequence exactly in GenBank, use either its GI or Accession.Version. To retrieve the most up-to-date sequence, use the accession number without version.



- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

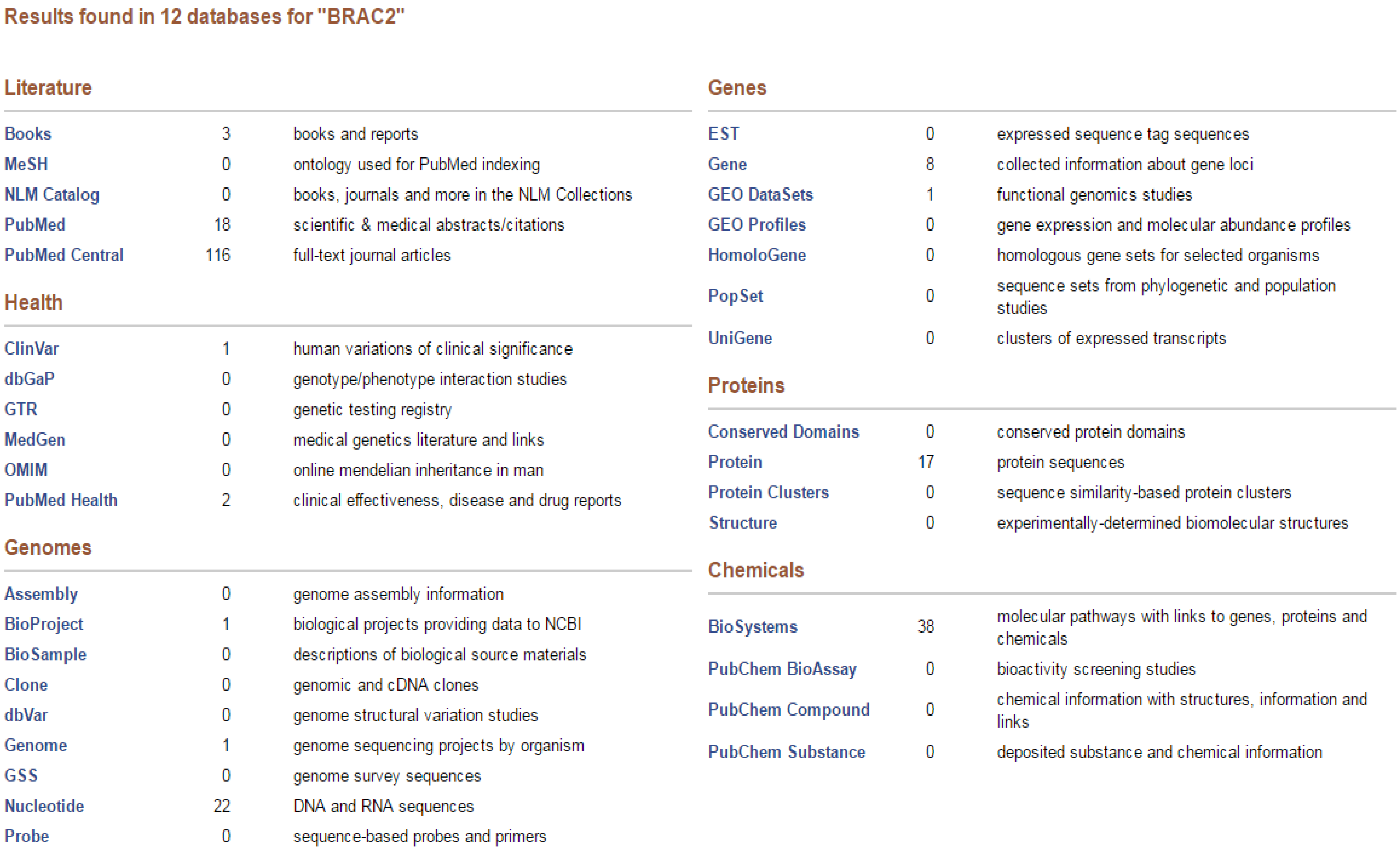
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

HTTPS at NCBI: Guidance for NCBI web API users
27 Jul 2016
As originally announced on June 10, NCBI will be moving all web services to

dbSNP build 148 for corn, fruit fly, rice and 8 other organisms available
26 Jul 2016
dbSNP build 148 is accessible on the web and via FTP. This release includes

August 3rd webinar: NCBI Targeted Loci: RefSeq Ribosomal RNA Sequences for Identification and Phylogenetic Analysis
21 Jul 2016
On August 3rd, NCBI staff will present a



Results found in 12 databases for "BRAC2"

Literature

Books	3	books and reports
MeSH	0	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	18	scientific & medical abstracts/citations
PubMed Central	116	full-text journal articles

Health

ClinVar	1	human variations of clinical significance
dbGaP	0	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	0	medical genetics literature and links
OMIM	0	online mendelian inheritance in man
PubMed Health	2	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	1	biological projects providing data to NCBI
BioSample	0	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	0	genome structural variation studies
Genome	1	genome sequencing projects by organism
GSS	0	genome survey sequences
Nucleotide	22	DNA and RNA sequences
Probe	0	sequence-based probes and primers

Genes

EST	0	expressed sequence tag sequences
Gene	8	collected information about gene loci
GEO DataSets	1	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	0	clusters of expressed transcripts

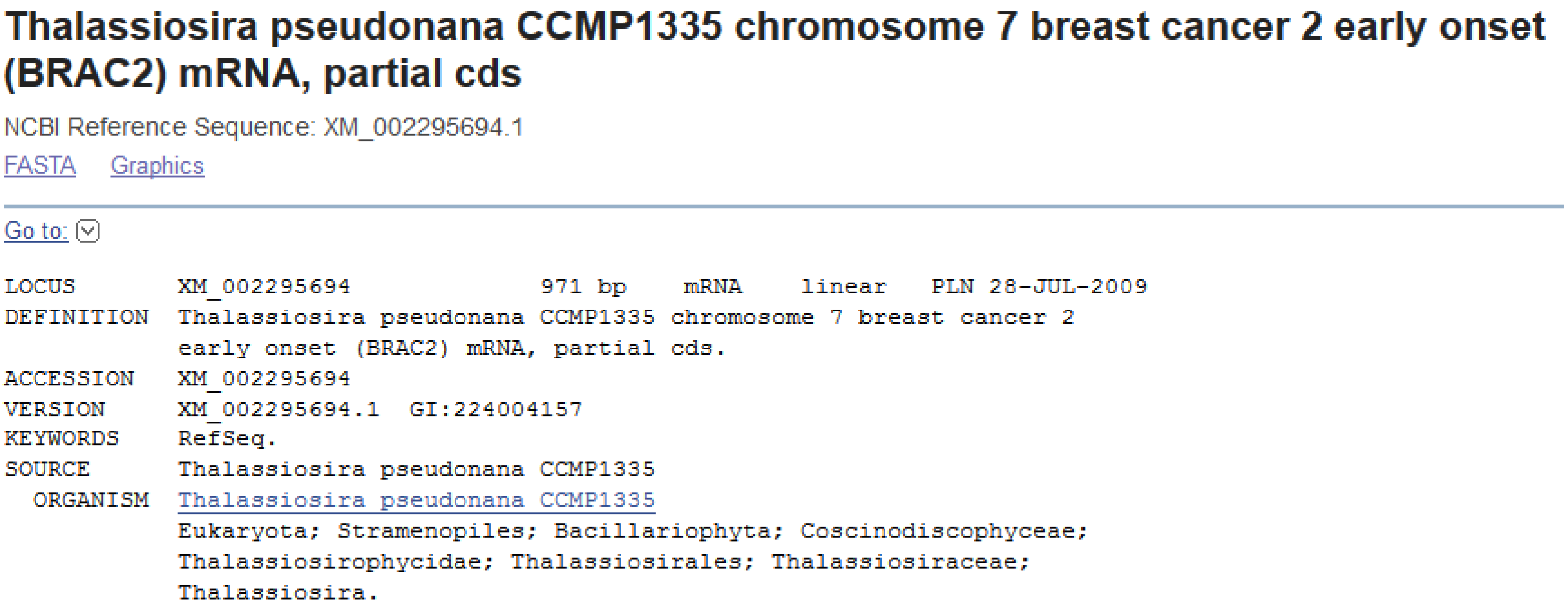
Proteins

Conserved Domains	0	conserved protein domains
Protein	17	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	0	experimentally-determined biomolecular structures

Chemicals

BioSystems	38	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	0	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	0	deposited substance and chemical information

GenBank Flatfile Format (GBFF)



Thalassiosira pseudonana CCMP1335 chromosome 7 breast cancer 2 early onset (BRAC2) mRNA, partial cds

NCBI Reference Sequence: XM_002295694.1

FASTA Graphics

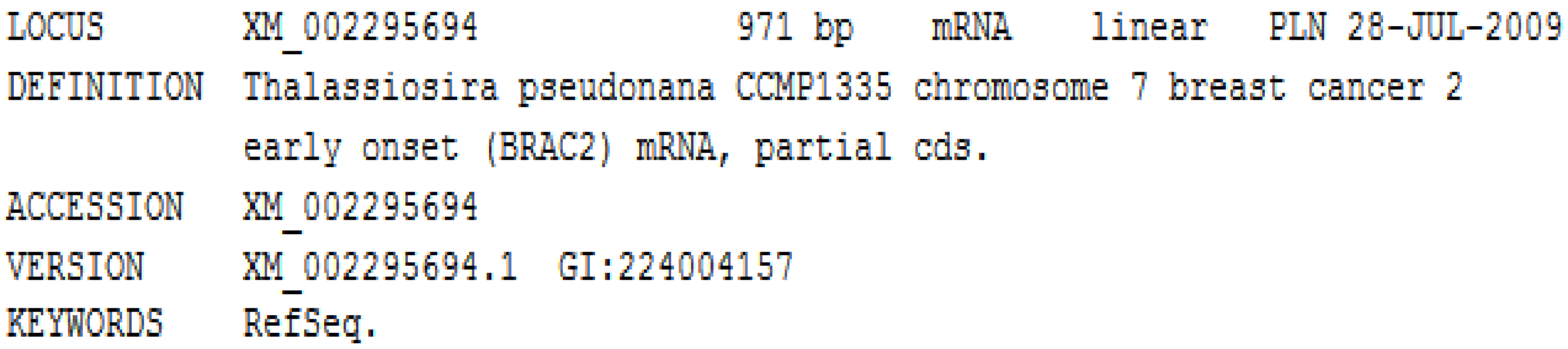
Go to: ⌄

```
LOCUS       XM_002295694             971 bp    mRNA    linear   PLN 28-JUL-2009
DEFINITION  Thalassiosira pseudonana CCMP1335 chromosome 7 breast cancer 2
            early onset (BRAC2) mRNA, partial cds.
ACCESSION   XM_002295694
VERSION     XM_002295694.1  GI:224004157
KEYWORDS    RefSeq.
SOURCE      Thalassiosira pseudonana CCMP1335
  ORGANISM  Thalassiosira pseudonana CCMP1335
            Eukaryota; Stramenopiles; Bacillariophyta; Coscinodiscophyceae;
            Thalassiosirophycidae; Thalassiosirales; Thalassiosiraceae;
            Thalassiosira.
```

- The GenBank flatfile format (GBFF) explain the nucleotide sequences of a specific gene. It contains all of the information associated with the sequence, as well as the sequence itself.
  The GBFF has 3 parts: the header, the features, and the sequence itself.

LOCUS	XM_002295694	971 bp	mRNA	linear	PLN 28-JUL-2009
	identifier	length	source	type	NCBI entry date taxonomic group

GenBank flatfile format - Header



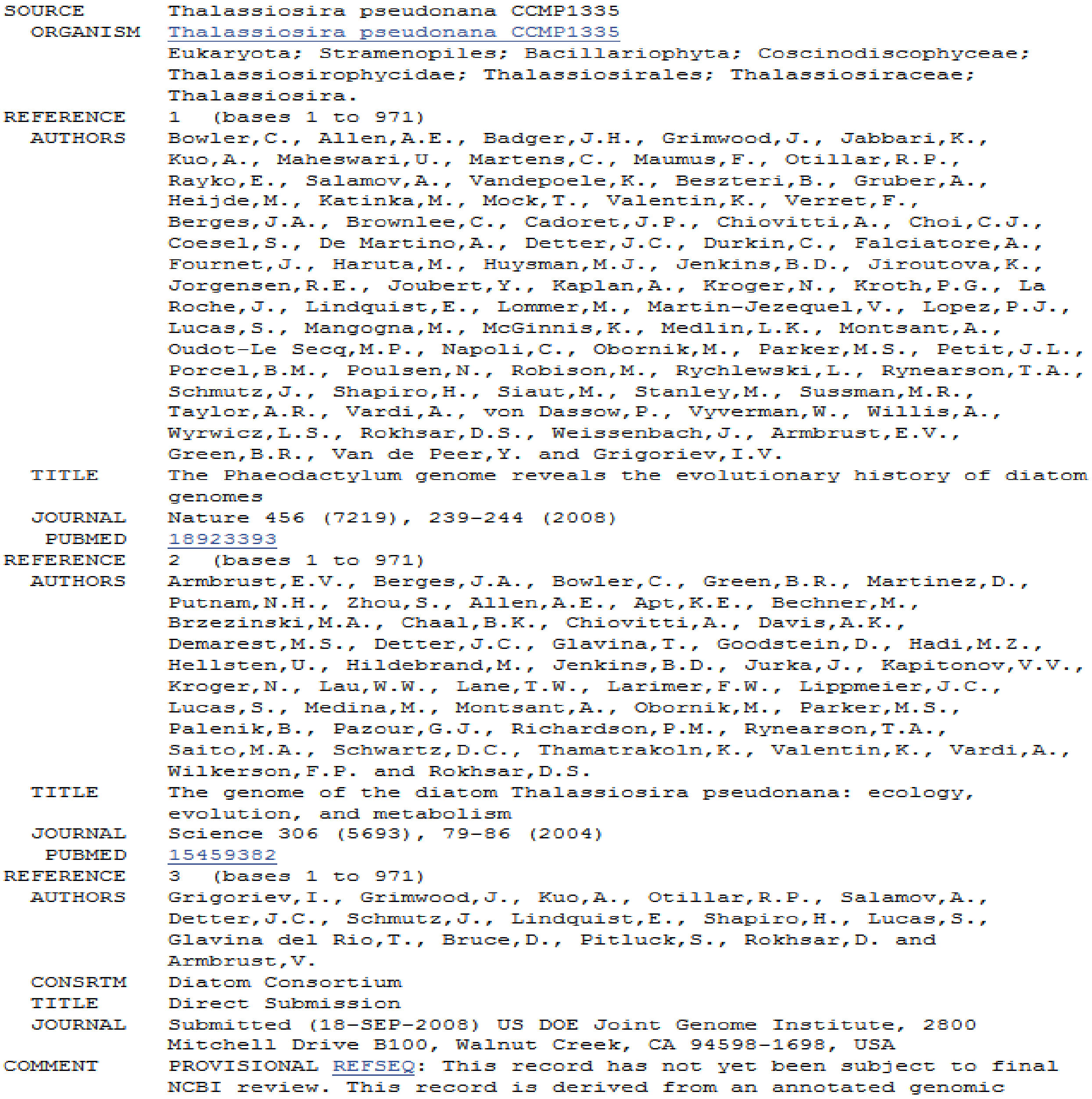
```
LOCUS       XM_002295694             971 bp    mRNA    linear   PLN 28-JUL-2009
DEFINITION  Thalassiosira pseudonana CCMP1335 chromosome 7 breast cancer 2
            early onset (BRAC2) mRNA, partial cds.
ACCESSION   XM_002295694
VERSION     XM_002295694.1  GI:224004157
KEYWORDS    RefSeq.
```

DEFINITION: The biology of the molecule in a sentence.

ACCESSION: Code(s)

VERSION: Number; GI number

KEYWORDS: Keywords as defined by the submitters



```
SOURCE      Thalassiosira pseudonana CCMP1335
  ORGANISM  Thalassiosira pseudonana CCMP1335
            Eukaryota; Stramenopiles; Bacillariophyta; Coscinodiscophyceae;
            Thalassiosirophycidae; Thalassiosirales; Thalassiosiraceae;
            Thalassiosira.
REFERENCE   1  (bases 1 to 971)
  AUTHORS   Bowler,C., Allen,A.E., Badger,J.H., Grimwood,J., Jabbari,K.,
            Kuo,A., Maheswari,U., Martens,C., Maumus,F., Otillar,R.P.,
            Rayko,E., Salamov,A., Vandepoele,K., Beszteri,B., Gruber,A.,
            Heijde,M., Katinka,M., Mock,T., Valentin,K., Verret,F.,
            Berges,J.A., Brownlee,C., Cadoret,J.P., Chiovitti,A., Choi,C.J.,
            Coesel,S., De Martino,A., Detter,J.C., Durkin,C., Falciatore,A.,
            Fournet,J., Haruta,M., Huysman,M.J., Jenkins,B.D., Jiroutova,K.,
            Jorgensen,R.E., Joubert,Y., Kaplan,A., Kroger,N., Kroth,P.G., La
            Roche,J., Lindquist,E., Lommer,M., Martin-Jezequel,V., Lopez,P.J.,
            Lucas,S., Mangogna,M., McGinnis,K., Medlin,L.K., Montsant,A.,
            Oudot-Le Secq,M.P., Napoli,C., Obornik,M., Parker,M.S., Petit,J.L.,
            Porcel,B.M., Poulsen,N., Robison,M., Rychlewski,L., Rynearson,T.A.,
            Schmutz,J., Shapiro,H., Siaut,M., Stanley,M., Sussman,M.R.,
            Taylor,A.R., Vardi,A., von Dassow,P., Vyverman,W., Willis,A.,
            Wyrwicz,L.S., Rokhsar,D.S., Weissenbach,J., Armbrust,E.V.,
            Green,B.R., Van de Peer,Y. and Grigoriev,I.V.
  TITLE     The Phaeodactylum genome reveals the evolutionary history of diatom
            genomes
  JOURNAL   Nature 456 (7219), 239-244 (2008)
   PUBMED   18923393
REFERENCE   2  (bases 1 to 971)
  AUTHORS   Armbrust,E.V., Berges,J.A., Bowler,C., Green,B.R., Martinez,D.,
            Putnam,N.H., Zhou,S., Allen,A.E., Apt,K.E., Bechner,M.,
            Brzezinski,M.A., Chaal,B.K., Chiovitti,A., Davis,A.K.,
            Demarest,M.S., Detter,J.C., Glavina,T., Goodstein,D., Hadi,M.Z.,
            Hellsten,U., Hildebrand,M., Jenkins,B.D., Jurka,J., Kapitonov,V.V.,
            Kroger,N., Lau,W.W., Lane,T.W., Larimer,F.W., Lippmeier,J.C.,
            Lucas,S., Medina,M., Montsant,A., Obornik,M., Parker,M.S.,
            Palenik,B., Pazour,G.J., Richardson,P.M., Rynearson,T.A.,
            Saito,M.A., Schwartz,D.C., Thamatrakoln,K., Valentin,K., Vardi,A.,
            Wilkerson,F.P. and Rokhsar,D.S.
  TITLE     The genome of the diatom Thalassiosira pseudonana: ecology,
            evolution, and metabolism
  JOURNAL   Science 306 (5693), 79-86 (2004)
   PUBMED   15459382
REFERENCE   3  (bases 1 to 971)
  AUTHORS   Grigoriev,I., Grimwood,J., Kuo,A., Otillar,R.P., Salamov,A.,
            Detter,J.C., Schmutz,J., Lindquist,E., Shapiro,H., Lucas,S.,
            Glavina del Rio,T., Bruce,D., Pitluck,S., Rokhsar,D. and
            Armbrust,V.
  CONSRTM   Diatom Consortium
  TITLE     Direct Submission
  JOURNAL   Submitted (18-SEP-2008) US DOE Joint Genome Institute, 2800
            Mitchell Drive B100, Walnut Creek, CA 94598-1698, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. This record is derived from an annotated genomic
```

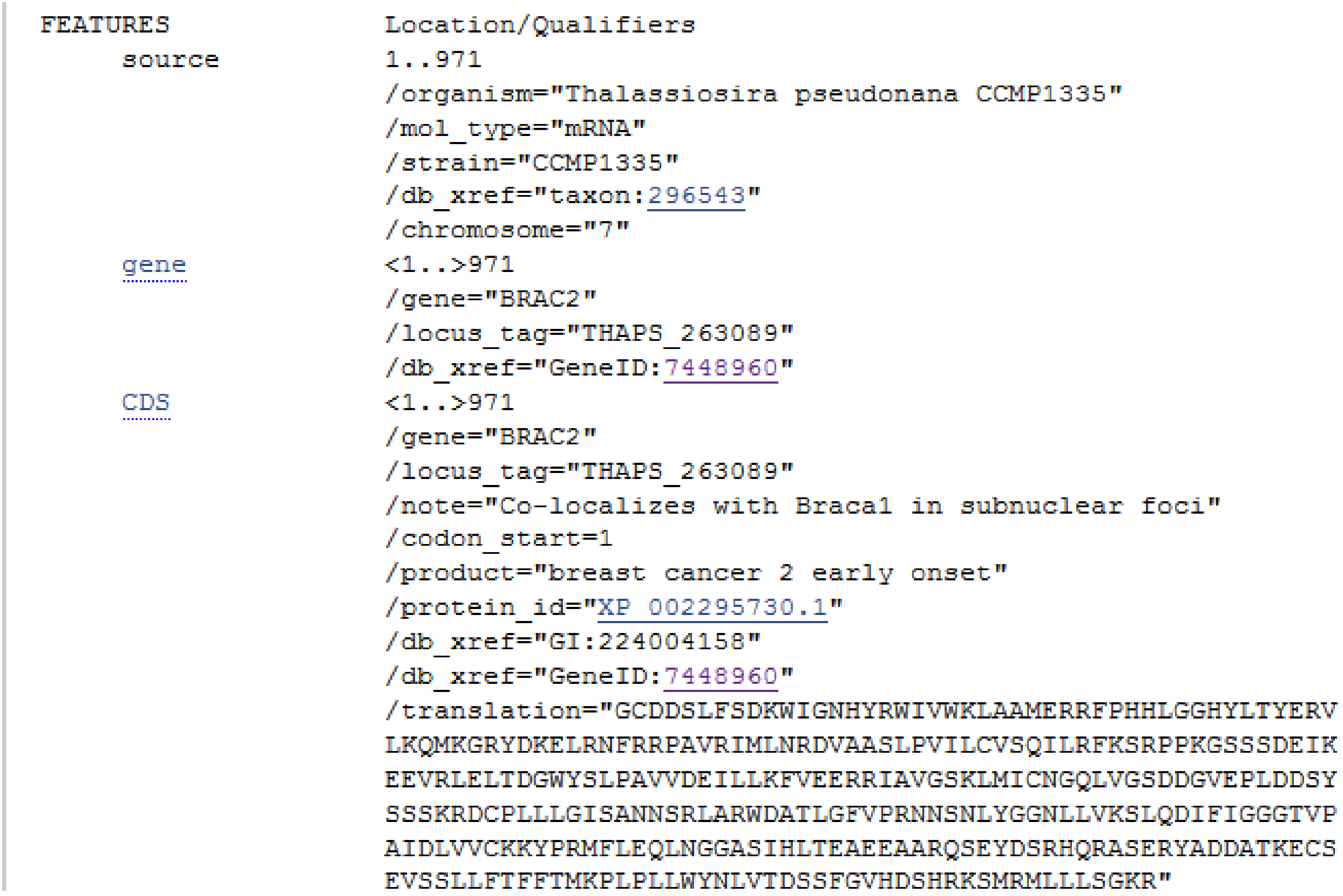
SOURCE: Contains organism name
ORGANISM: Contains complete taxonomic information from the NCBI taxonomy server.

REFERENCE: Details on a publication about the sequence.
COMMENT: Contains misc. information and revision details.

GenBank Flatfile Format – Features

A direct representation of the biological information in the record.

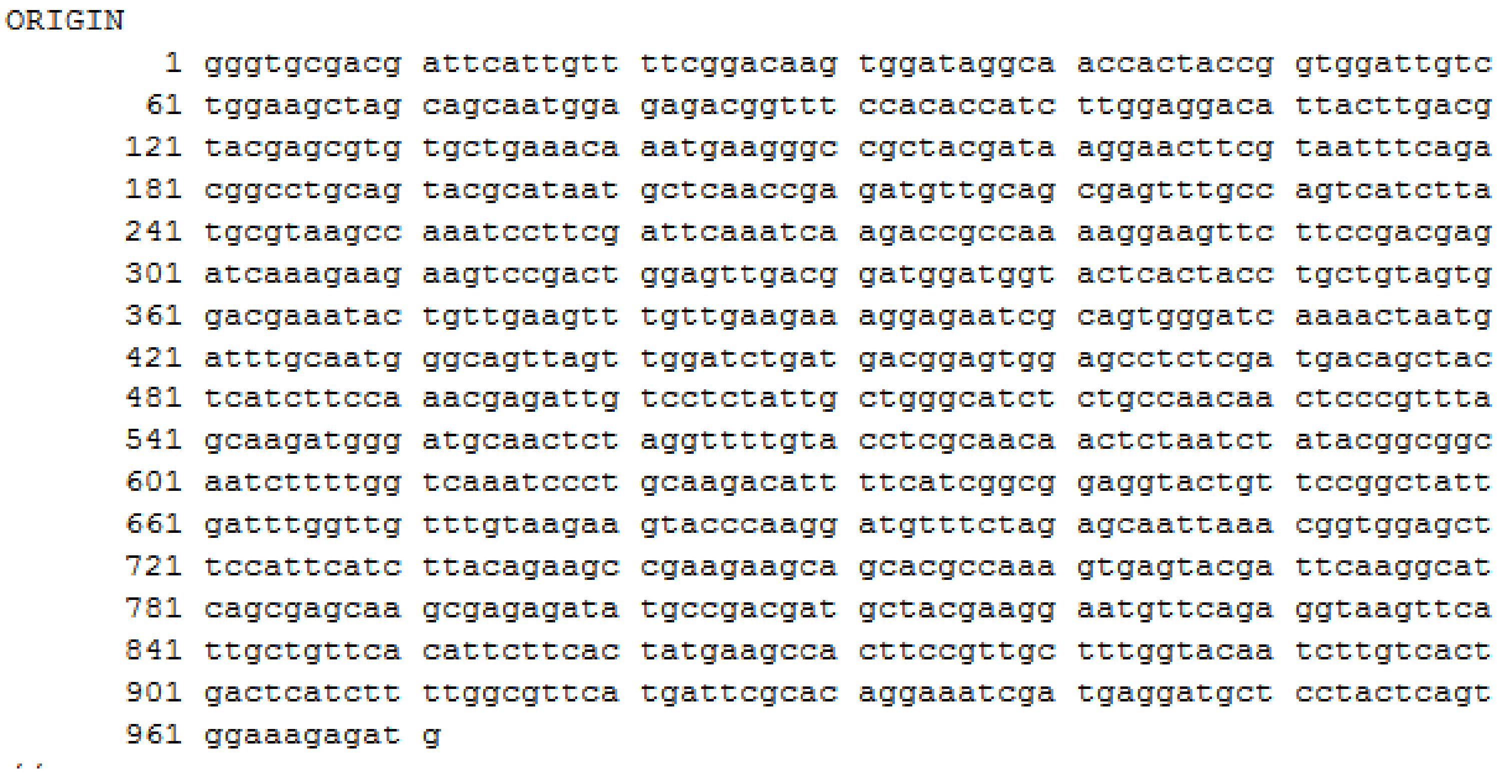
- The Source Feature must be present in all GenBank records, and contains information as to where the molecule comes from /organism = "Homo sapiens", and, potentially, map, chromosome and tissue type information.
- In some records the CDS (coding sequence) feature is present:



```
FEATURES             Location/Qualifiers
     source          1..971
                     /organism="Thalassiosira pseudonana CCMP1335"
                     /mol_type="mRNA"
                     /strain="CCMP1335"
                     /db_xref="taxon:296543"
                     /chromosome="7"
     gene            <1..>971
                     /gene="BRAC2"
                     /locus_tag="THAPS_263089"
                     /db_xref="GeneID:7448960"
     CDS             <1..>971
                     /gene="BRAC2"
                     /locus_tag="THAPS_263089"
                     /note="Co-localizes with Braca1 in subnuclear foci"
                     /codon_start=1
                     /product="breast cancer 2 early onset"
                     /protein_id="XP_002295730.1"
                     /db_xref="GI:224004158"
                     /db_xref="GeneID:7448960"
                     /translation="GCDDSLFSDKWIGNHYRWIVWKLAAMERRFPHHLGGHYLTYERV
                     LKQMKGRYDKELRNFRRPAVRIMLNRDVAASLPVILCVSQILRFKSRPPKGSSSDEIK
                     EEVRLELTDGWYSLPAVVDEILLKFVEERRIAVGSKLMICNGQLVGSDDGVEPLDDSY
                     SSSKRDCPLLLGISANNSRLARWDATLGFVPRNNSNLYGGNLLVKSLQDIFIGGGTVP
                     AIDLVVCKKYPRMFLEQLNGGASIHLTEAEEAARQSEYDSRHQRASERYADDATKECS
                     EVSSLLFTFFTMKPLPLLWYNLVTDSSFGVHDSHRKSMRMLLLSGKR"
```

GenBank Flatfile Format – Sequence

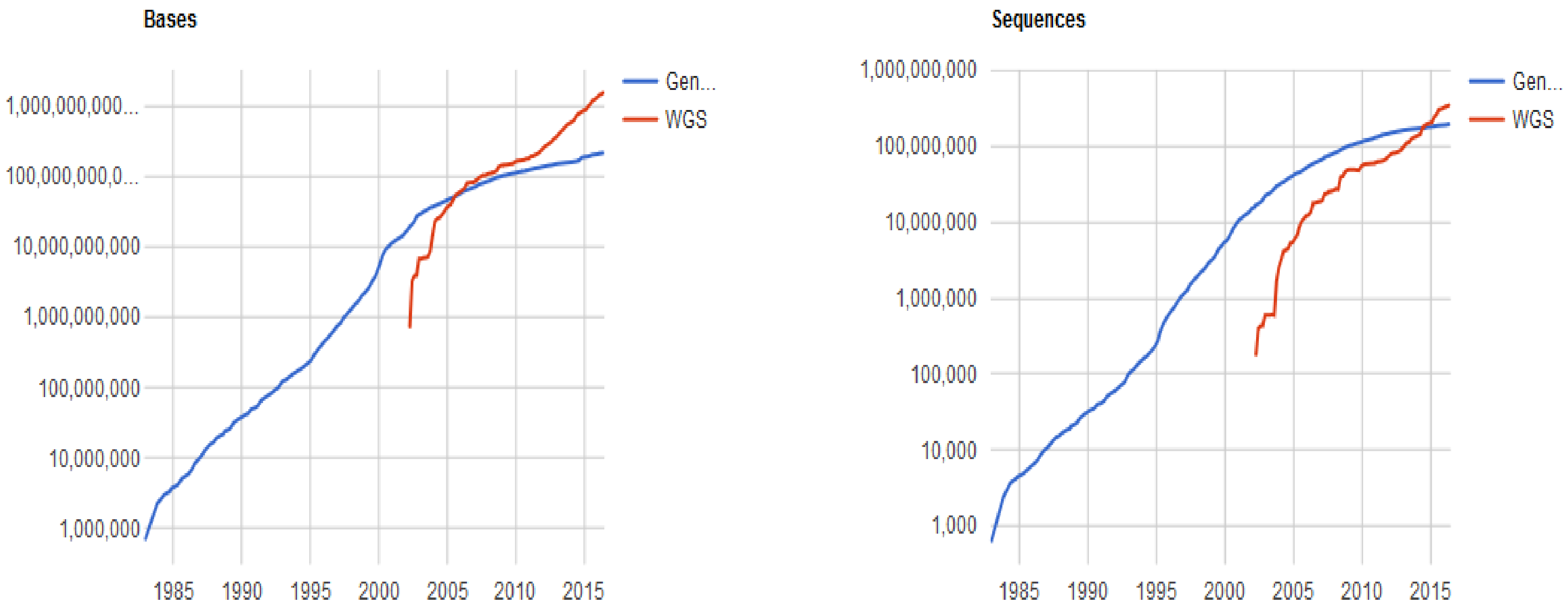
- The last part of the GenBank flat file record is the sequence itself:



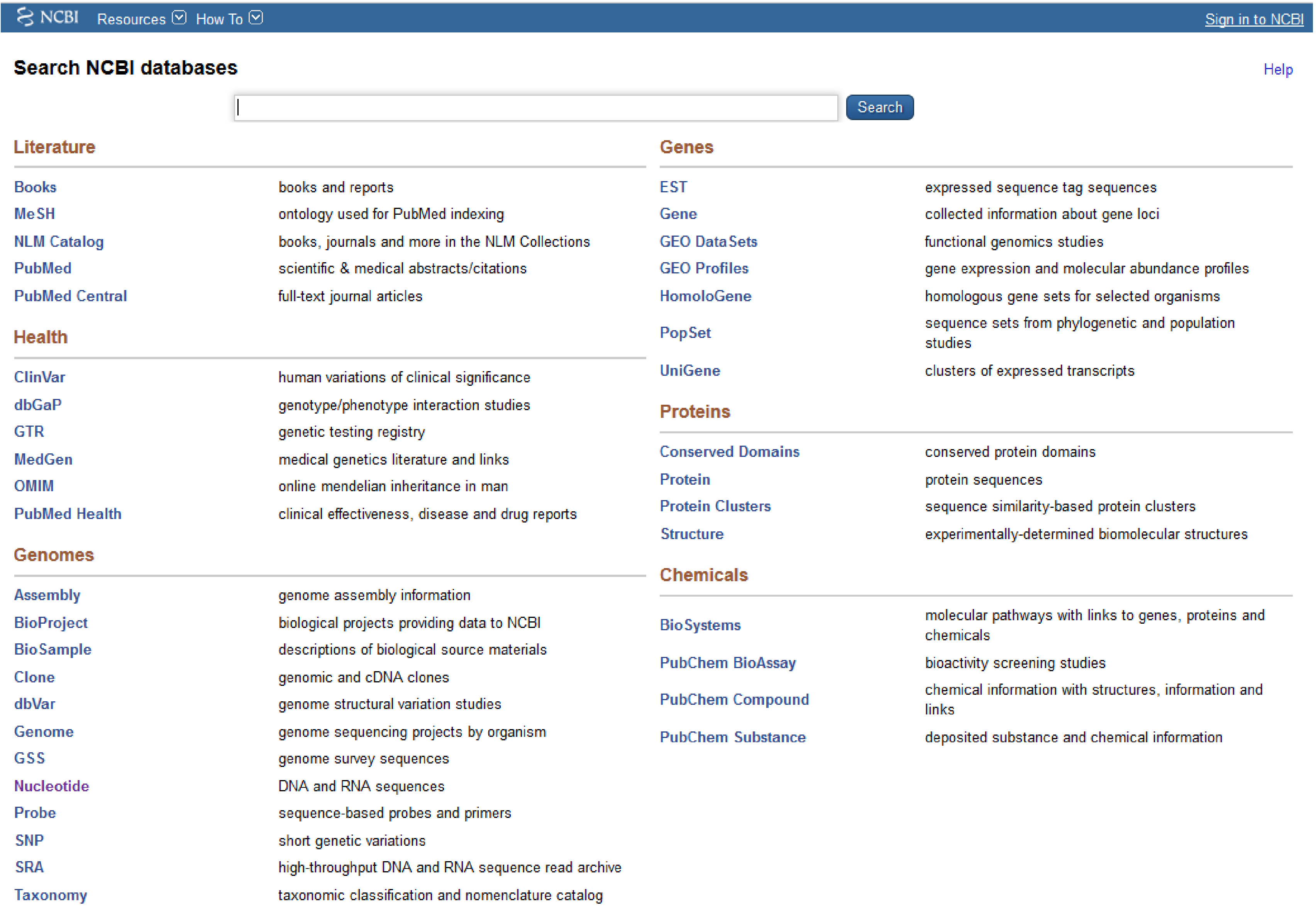
```
ORIGIN
        1 gggtgcgacg attcattgtt ttcggacaag tggataggca accactaccg gtggattgtc
       61 tggaagctag cagcaatgga gagacggttt ccacaccatc ttggaggaca ttacttgacg
      121 tacgagcgtg tgctgaaaca aatgaagggc cgctacgata aggaacttcg taatttcaga
      181 cggcctgcag tacgcataat gctcaaccga gatgttgcag cgagtttgcc agtcatctta
      241 tgcgtaagcc aaatccttcg attcaaatca agaccgccaa aaggaagttc ttccgacgag
      301 atcaaagaag aagtccgact ggagttgacg gatggatggt actcactacc tgctgtagtg
      361 gacgaaatac tgttgaagtt tgttgaagaa aggagaatcg cagtgggatc aaaactaatg
      421 atttgcaatg ggcagttagt tggatctgat gacggagtgg agcctctcga tgacagctac
      481 tcatcttcca aacgagattg tcctctattg ctgggcatct ctgccaacaa ctcccgttta
      541 gcaagatggg atgcaactct aggttttgta cctcgcaaca actctaatct atacggcggc
      601 aatcttttgg tcaaatccct gcaagacatt ttcatcggcg gaggtactgt tccggctatt
      661 gatttggttg tttgtaagaa gtacccaagg atgtttctag agcaattaaa cggtggagct
      721 tccattcatc ttacagaagc cgaagaagca gcacgccaaa gtgagtacga ttcaaggcat
      781 cagcgagcaa gcgagagata tgccgacgat gctacgaagg aatgttcaga ggtaagttca
      841 ttgctgttca cattcttcac tatgaagcca cttccgttgc tttggtacaa tcttgtcact
      901 gactcatctt ttggcgttca tgattcgcac aggaaatcga tgaggatgct cctactcagt
      961 ggaaagagat g
//
```

Nucleotide Databases – Growth of GenBank

GenBank and WGS Statistics



Other facilities in NCBI database



Search NCBI databases

Literature

Books	books and reports
MeSH	ontology used for PubMed indexing
NLM Catalog	books, journals and more in the NLM Collections
PubMed	scientific & medical abstracts/citations
PubMed Central	full-text journal articles

Health

ClinVar	human variations of clinical significance
dbGaP	genotype/phenotype interaction studies
GTR	genetic testing registry
MedGen	medical genetics literature and links
OMIM	online mendelian inheritance in man
PubMed Health	clinical effectiveness, disease and drug reports

Genomes

Assembly	genome assembly information
BioProject	biological projects providing data to NCBI
BioSample	descriptions of biological source materials
Clone	genomic and cDNA clones
dbVar	genome structural variation studies
Genome	genome sequencing projects by organism
GSS	genome survey sequences
Nucleotide	DNA and RNA sequences
Probe	sequence-based probes and primers
SNP	short genetic variations
SRA	high-throughput DNA and RNA sequence read archive
Taxonomy	taxonomic classification and nomenclature catalog

Genes

EST	expressed sequence tag sequences
Gene	collected information about gene loci
GEO DataSets	functional genomics studies
GEO Profiles	gene expression and molecular abundance profiles
HomoloGene	homologous gene sets for selected organisms
PopSet	sequence sets from phylogenetic and population studies
UniGene	clusters of expressed transcripts

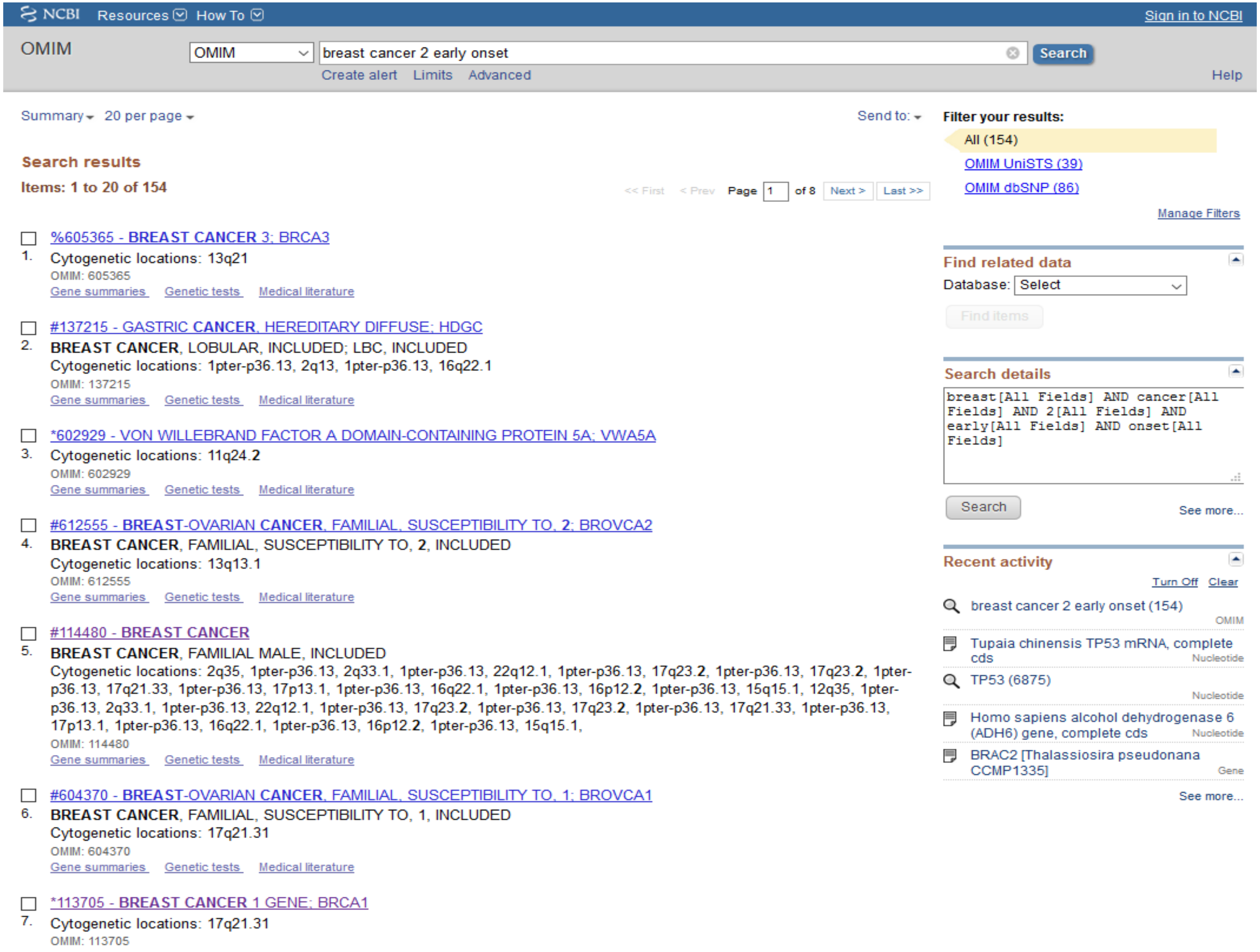
Proteins

Conserved Domains	conserved protein domains
Protein	protein sequences
Protein Clusters	sequence similarity-based protein clusters
Structure	experimentally-determined biomolecular structures

Chemicals

BioSystems	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	bioactivity screening studies
PubChem Compound	chemical information with structures, information and links
PubChem Substance	deposited substance and chemical information

Disease details



breast cancer 2 early onset

Search results

Items: 1 to 20 of 154

□ 1. %605365 - BREAST CANCER 3; BRCA3
Cytogenetic locations: 13q21
OMIM: 605365
Gene summaries Genetic tests Medical literature

□ 2. #137215 - GASTRIC CANCER, HEREDITARY DIFFUSE; HDGC
BREAST CANCER, LOBULAR, INCLUDED; LBC, INCLUDED
Cytogenetic locations: 1pter-p36.13, 2q13, 1pter-p36.13, 16q22.1
OMIM: 137215
Gene summaries Genetic tests Medical literature

□ 3. *602929 - VON WILLEBRAND FACTOR A DOMAIN-CONTAINING PROTEIN 5A; VWA5A
Cytogenetic locations: 11q24.2
OMIM: 602929
Gene summaries Genetic tests Medical literature

□ 4. #612555 - BREAST-OVARIAN CANCER, FAMILIAL, SUSCEPTIBILITY TO, 2; BROVCA2
BREAST CANCER, FAMILIAL, SUSCEPTIBILITY TO, 2, INCLUDED
Cytogenetic locations: 13q13.1
OMIM: 612555
Gene summaries Genetic tests Medical literature

□ 5. #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

□ 6. #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

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

Filter your results:

All (154)
OMIM UniSTS (39)
OMIM dbSNP (86)
Manage Filters

Find related data

Database: Select

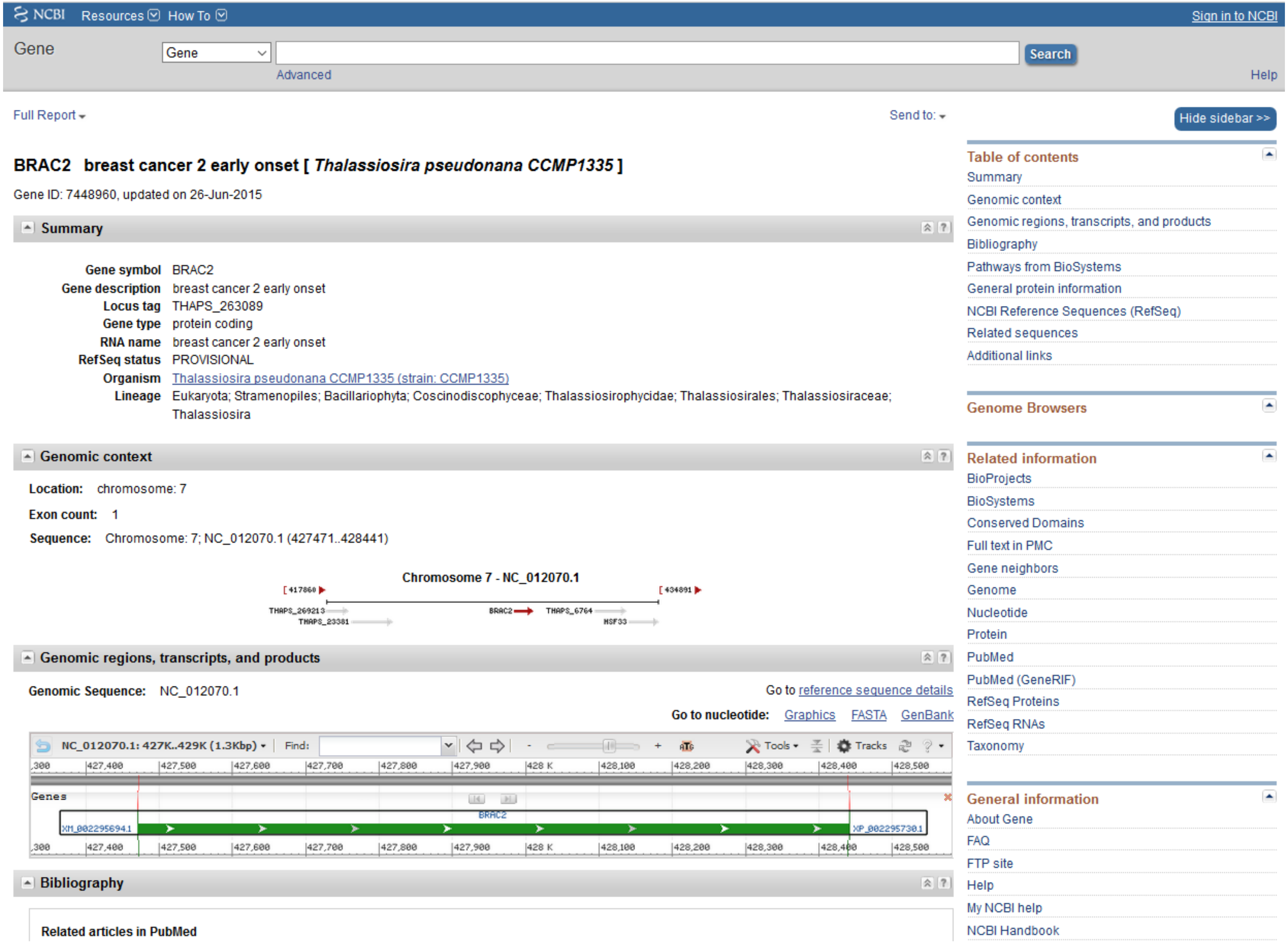
Search details

```
breast[All Fields] AND cancer[All
Fields] AND 2[All Fields] AND
early[All Fields] AND onset[All
Fields]
```

Recent activity

breast cancer 2 early onset (154) OMIM
Tupaia chinensis TP53 mRNA, complete cds Nucleotide
TP53 (6875) Nucleotide
Homo sapiens alcohol dehydrogenase 6 (ADH6) gene, complete cds Nucleotide
BRAC2 [Thalassiosira pseudonana CCMP1335] Gene

Gene Details



NCBI Resources How To Sign in to NCBI

Gene | Gene | Search | Help

Advanced

Full Report Send to: Hide sidebar >>

BRAC2 breast cancer 2 early onset [*Thalassiosira pseudonana CCMP1335*]

Gene ID: 7448960, updated on 26-Jun-2015

Summary

Gene symbol	BRAC2
Gene description	breast cancer 2 early onset
Locus tag	THAPS_263089
Gene type	protein coding
RNA name	breast cancer 2 early onset
RefSeq status	PROVISIONAL
Organism	Thalassiosira pseudonana CCMP1335 (strain: CCMP1335)
Lineage	Eukaryota; Stramenopiles; Bacillariophyta; Coscinodiscophyceae; Thalassiosirophycidae; Thalassiosirales; Thalassiosiraceae; Thalassiosira

Genomic context

Location: chromosome: 7

Exon count: 1

Sequence: Chromosome: 7; NC_012070.1 (427471..428441)

Chromosome 7 - NC_012070.1

Genomic regions, transcripts, and products

Genomic Sequence: NC_012070.1

Go to reference sequence details

Go to nucleotide: Graphics FASTA GenBank

Bibliography

Related articles in PubMed

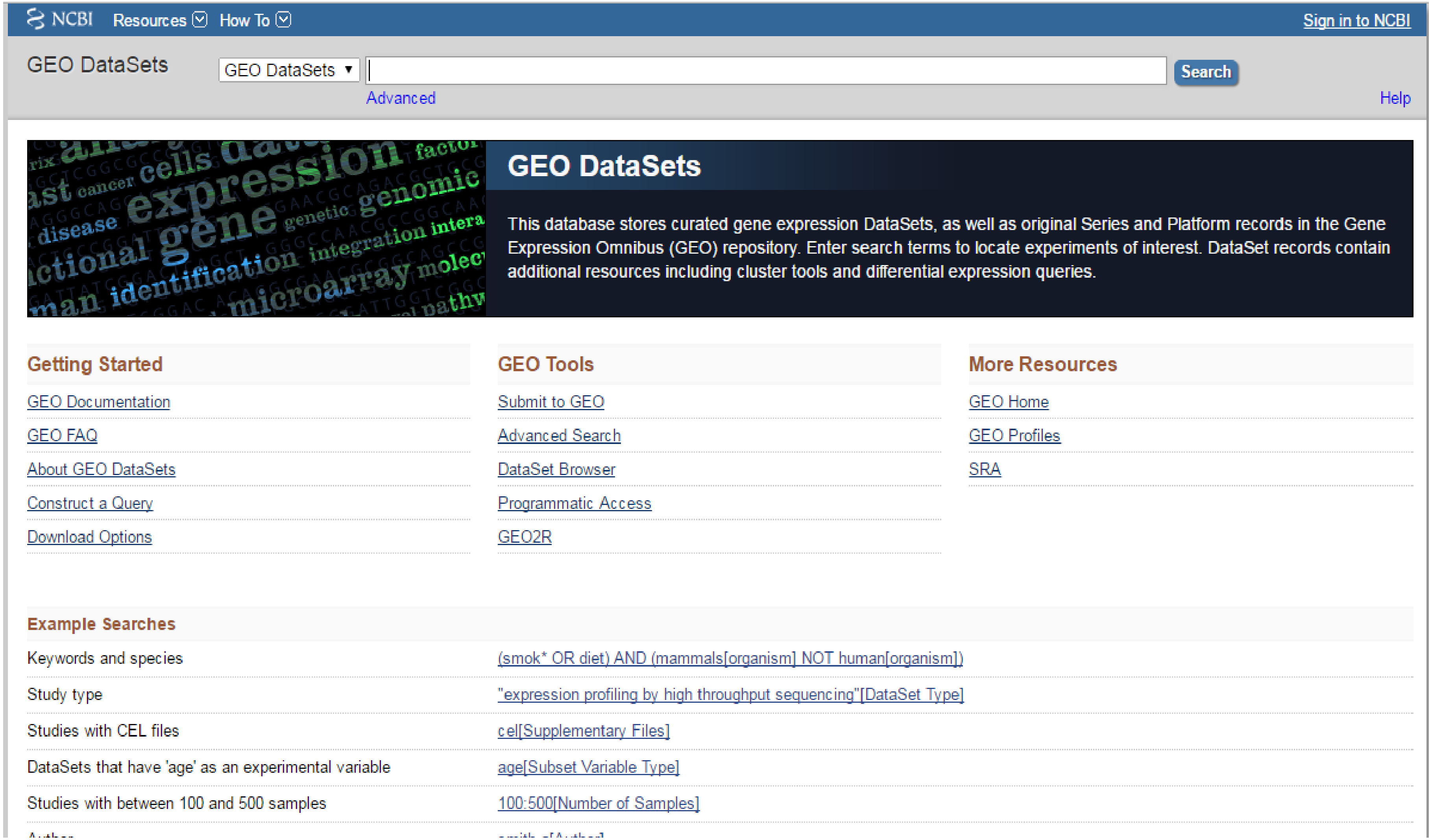
Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Pathways from BioSystems
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Genome Browsers

Related information
- BioProjects
- BioSystems
- Conserved Domains
- Full text in PMC
- Gene neighbors
- Genome
- Nucleotide
- Protein
- PubMed
- PubMed (GeneRIF)
- RefSeq Proteins
- RefSeq RNAs
- Taxonomy

General information
- About Gene
- FAQ
- FTP site
- Help
- My NCBI help
- NCBI Handbook

Gene expression details....



GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Getting Started

- GEO Documentation
- GEO FAQ
- About GEO DataSets
- Construct a Query
- Download Options

GEO Tools

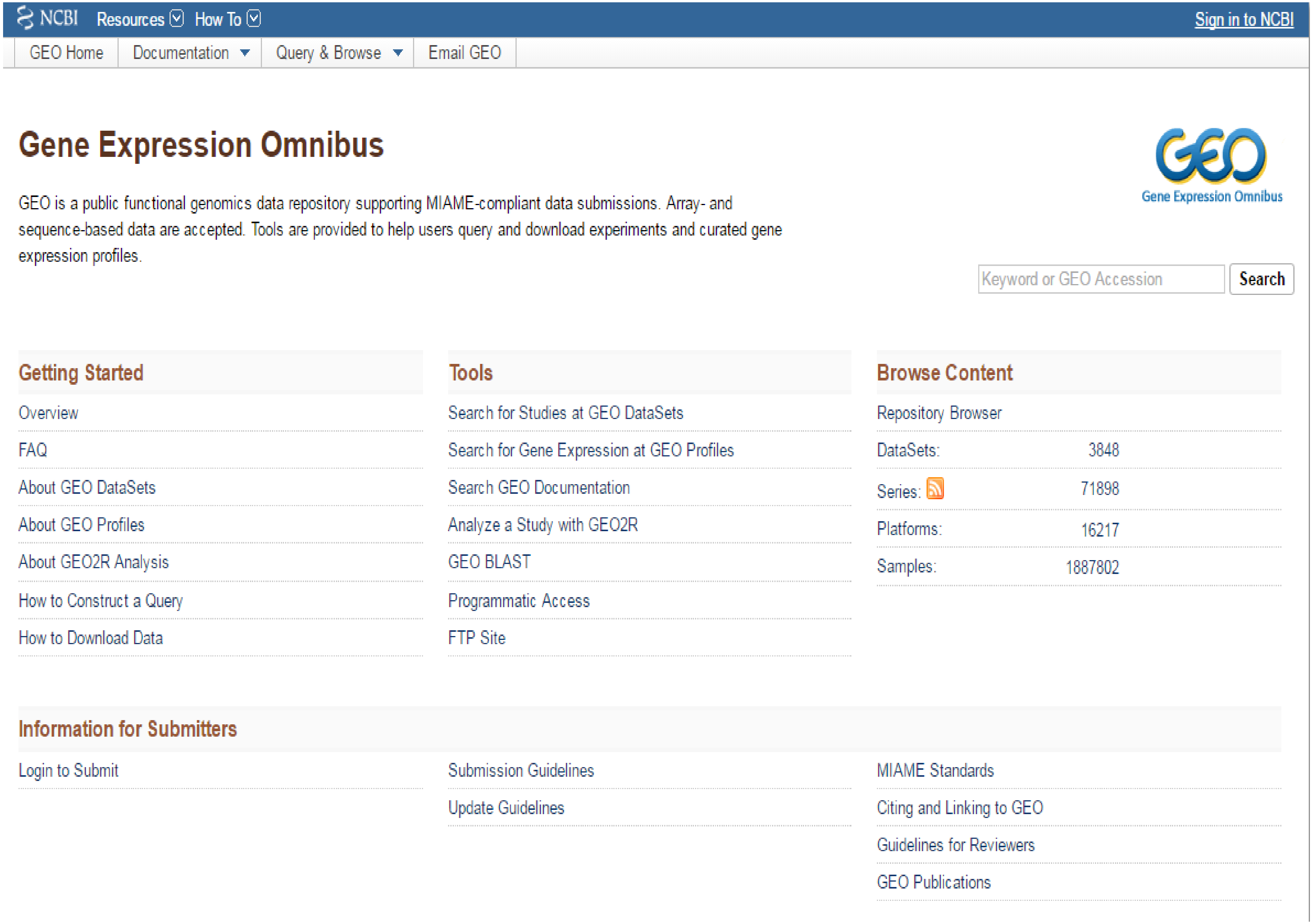
- Submit to GEO
- Advanced Search
- DataSet Browser
- Programmatic Access
- GEO2R

More Resources

- GEO Home
- GEO Profiles
- SRA

Example Searches

Keywords and species	(smok* OR diet) AND (mammals[organism] NOT human[organism])
Study type	"expression profiling by high throughput sequencing"[DataSet Type]
Studies with CEL files	cel[Supplementary Files]
DataSets that have 'age' as an experimental variable	age[Subset Variable Type]
Studies with between 100 and 500 samples	100:500[Number of Samples]
Author	smith c[Author]



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

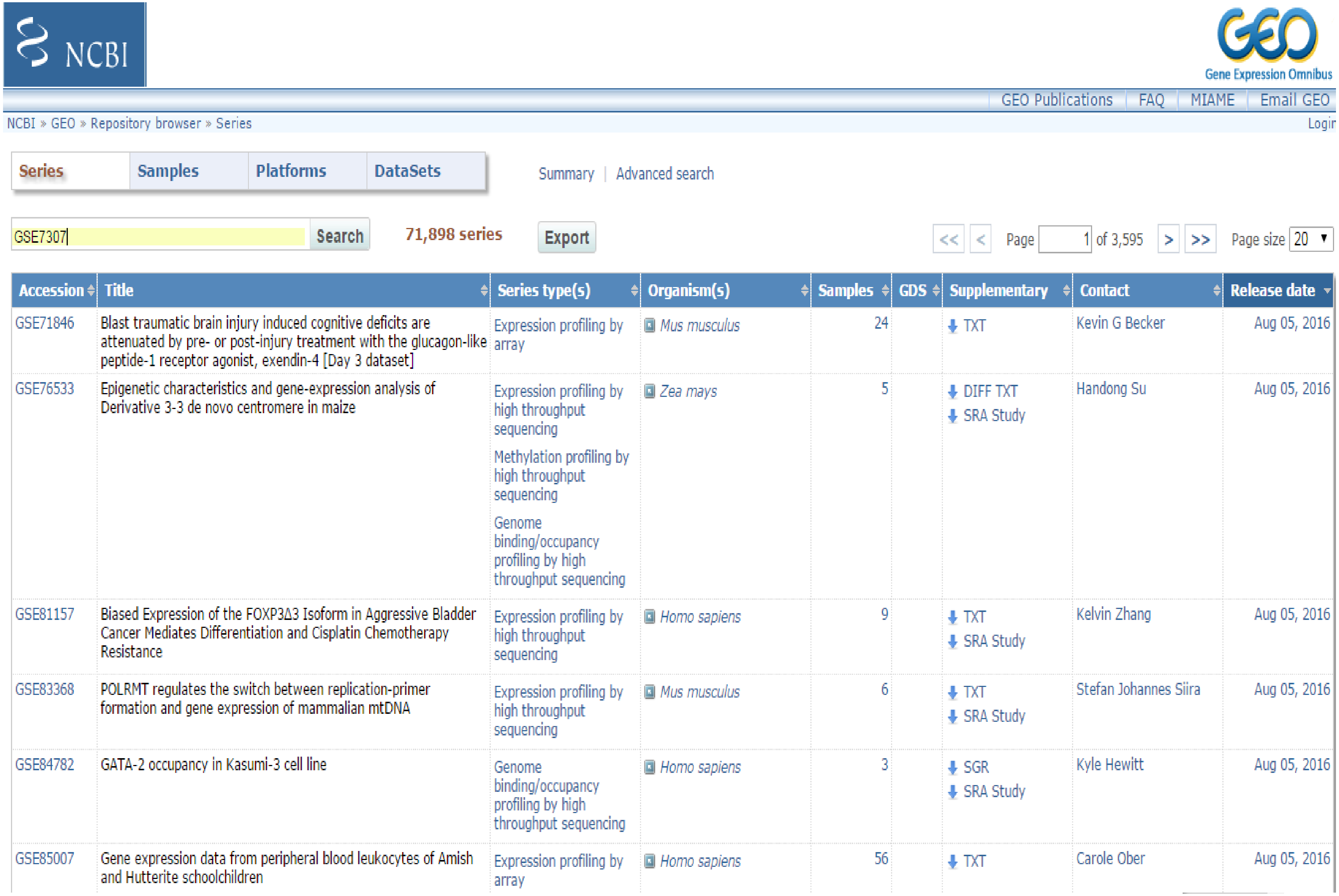
Browse Content

Repository Browser

DataSets:	3848
Series:	71898
Platforms:	16217
Samples:	1887802

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications



NCBI
GEO
Gene Expression Omnibus
GEO Publications | FAQ | MIAME | Email GEO
NCBI » GEO » Repository browser » Series
Login
Series
Samples
Platforms
DataSets
Summary | Advanced search

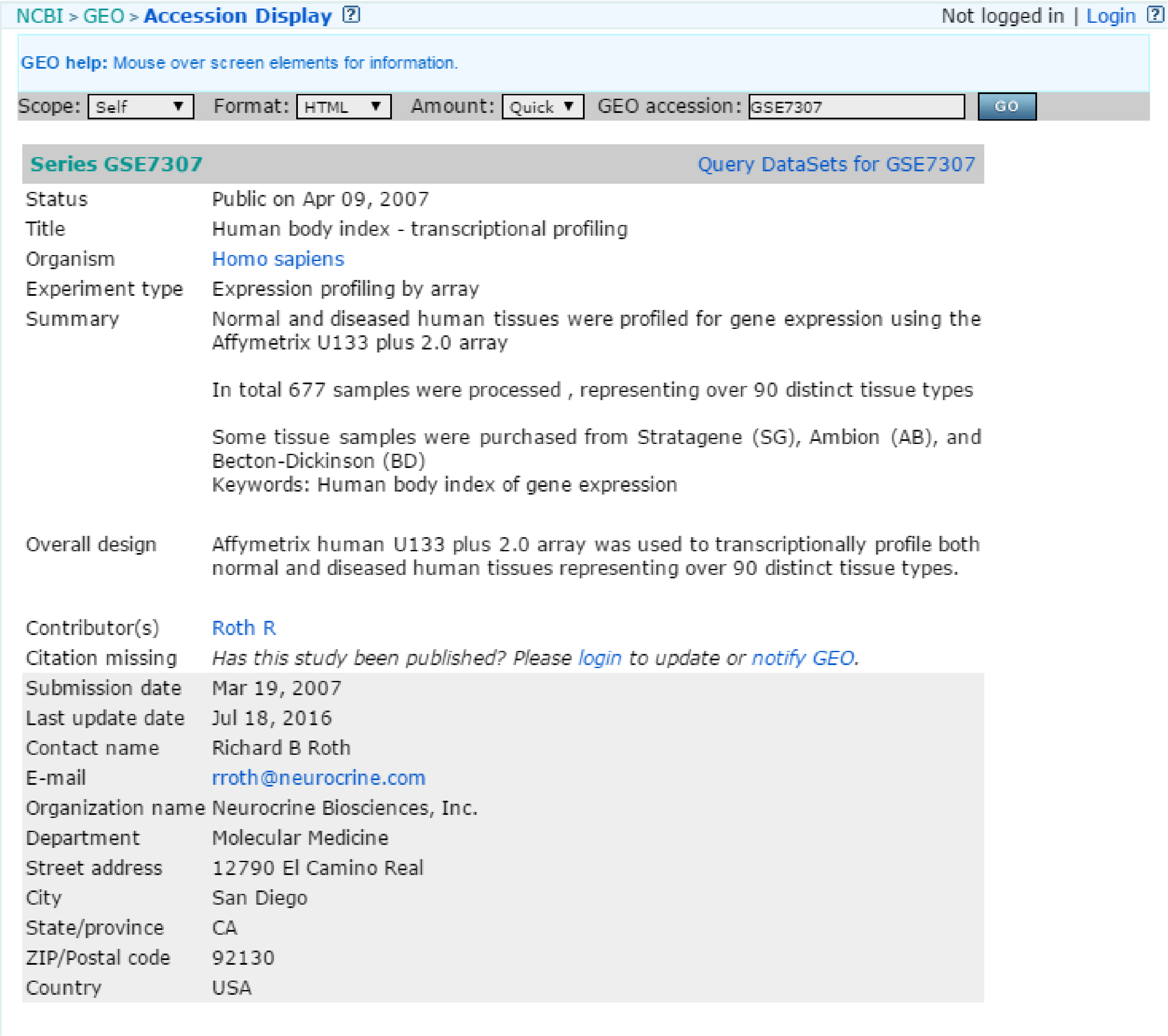
GSE7307| Search

71,898 series

Export

<< < Page 1 of 3,595 > >> Page size 20

Accession	Title	Series type(s)	Organism(s)	Samples	GDS	Supplementary	Contact	Release date
GSE71846	Blast traumatic brain injury induced cognitive deficits are attenuated by pre- or post-injury treatment with the glucagon-like peptide-1 receptor agonist, exendin-4 [Day 3 dataset]	Expression profiling by array	*Mus musculus*	24		↓ TXT	Kevin G Becker	Aug 05, 2016
GSE76533	Epigenetic characteristics and gene-expression analysis of Derivative 3-3 de novo centromere in maize	Expression profiling by high throughput sequencing Methylation profiling by high throughput sequencing Genome binding/occupancy profiling by high throughput sequencing	*Zea mays*	5		↓ DIFF TXT ↓ SRA Study	Handong Su	Aug 05, 2016
GSE81157	Biased Expression of the FOXP3Δ3 Isoform in Aggressive Bladder Cancer Mediates Differentiation and Cisplatin Chemotherapy Resistance	Expression profiling by high throughput sequencing	*Homo sapiens*	9		↓ TXT ↓ SRA Study	Kelvin Zhang	Aug 05, 2016
GSE83368	POLRMT regulates the switch between replication-primer formation and gene expression of mammalian mtDNA	Expression profiling by high throughput sequencing	*Mus musculus*	6		↓ TXT ↓ SRA Study	Stefan Johannes Siira	Aug 05, 2016
GSE84782	GATA-2 occupancy in Kasumi-3 cell line	Genome binding/occupancy profiling by high throughput sequencing	*Homo sapiens*	3		↓ SGR ↓ SRA Study	Kyle Hewitt	Aug 05, 2016
GSE85007	Gene expression data from peripheral blood leukocytes of Amish and Hutterite schoolchildren	Expression profiling by array	*Homo sapiens*	56		↓ TXT	Carole Ober	Aug 05, 2016

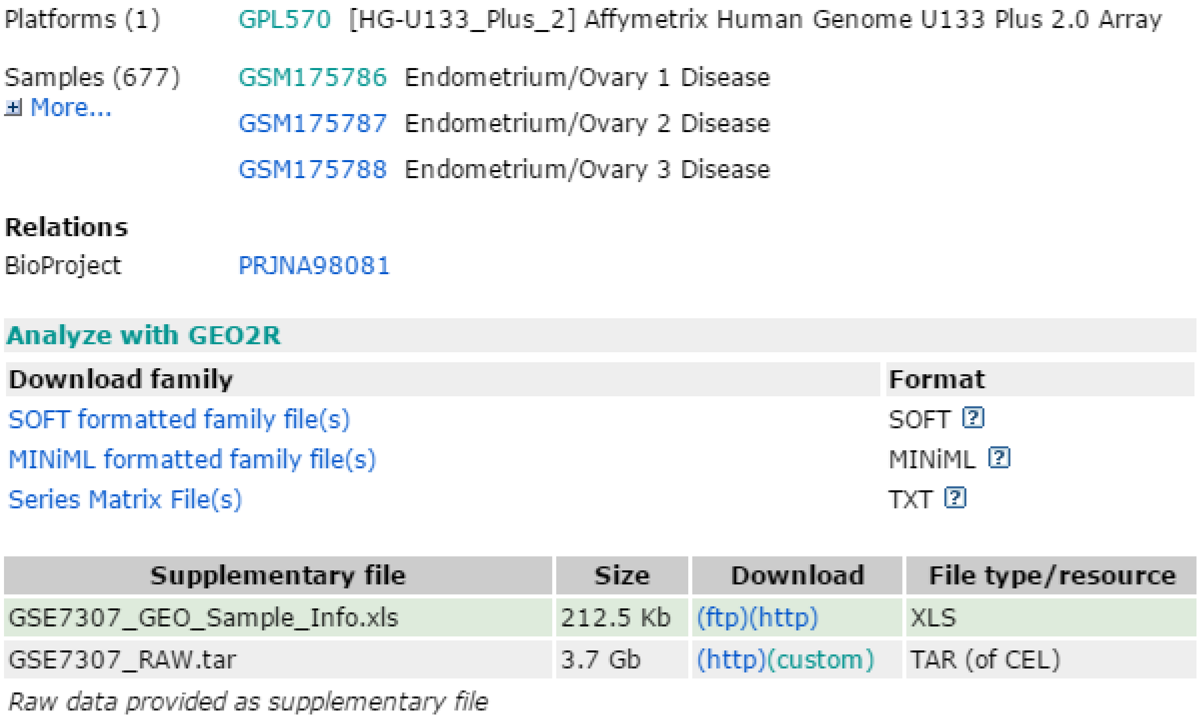


GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE7307

Series GSE7307 Query DataSets for GSE7307

Status	Public on Apr 09, 2007
Title	Human body index - transcriptional profiling
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Normal and diseased human tissues were profiled for gene expression using the Affymetrix U133 plus 2.0 array In total 677 samples were processed , representing over 90 distinct tissue types Some tissue samples were purchased from Stratagene (SG), Ambion (AB), and Becton-Dickinson (BD) Keywords: Human body index of gene expression
Overall design	Affymetrix human U133 plus 2.0 array was used to transcriptionally profile both normal and diseased human tissues representing over 90 distinct tissue types.
Contributor(s)	Roth R
Citation missing	*Has this study been published? Please login to update or notify GEO.*
Submission date	Mar 19, 2007
Last update date	Jul 18, 2016
Contact name	Richard B Roth
E-mail	rroth@neurocrine.com
Organization name	Neurocrine Biosciences, Inc.
Department	Molecular Medicine
Street address	12790 El Camino Real
City	San Diego
State/province	CA
ZIP/Postal code	92130
Country	USA



Platforms (1)	GPL570	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (677) ⊞ More...	GSM175786	Endometrium/Ovary 1 Disease
	GSM175787	Endometrium/Ovary 2 Disease
	GSM175788	Endometrium/Ovary 3 Disease

Relations

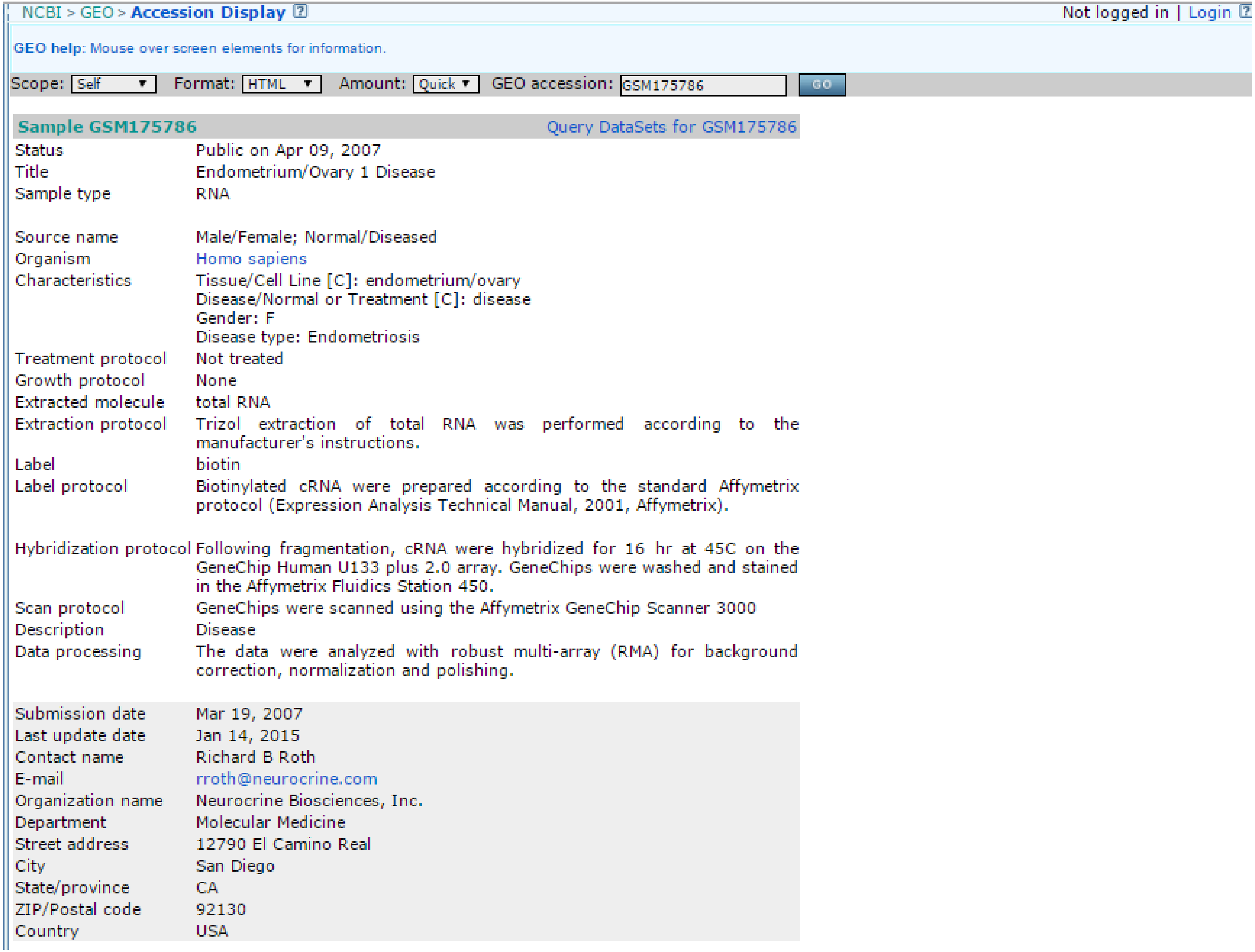
BioProject	PRJNA98081

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Supplementary file	Size	Download	File type/resource
GSE7307_GEO_Sample_Info.xls	212.5 Kb	(ftp)(http)	XLS
GSE7307_RAW.tar	3.7 Gb	(http)(custom)	TAR (of CEL)

Raw data provided as supplementary file



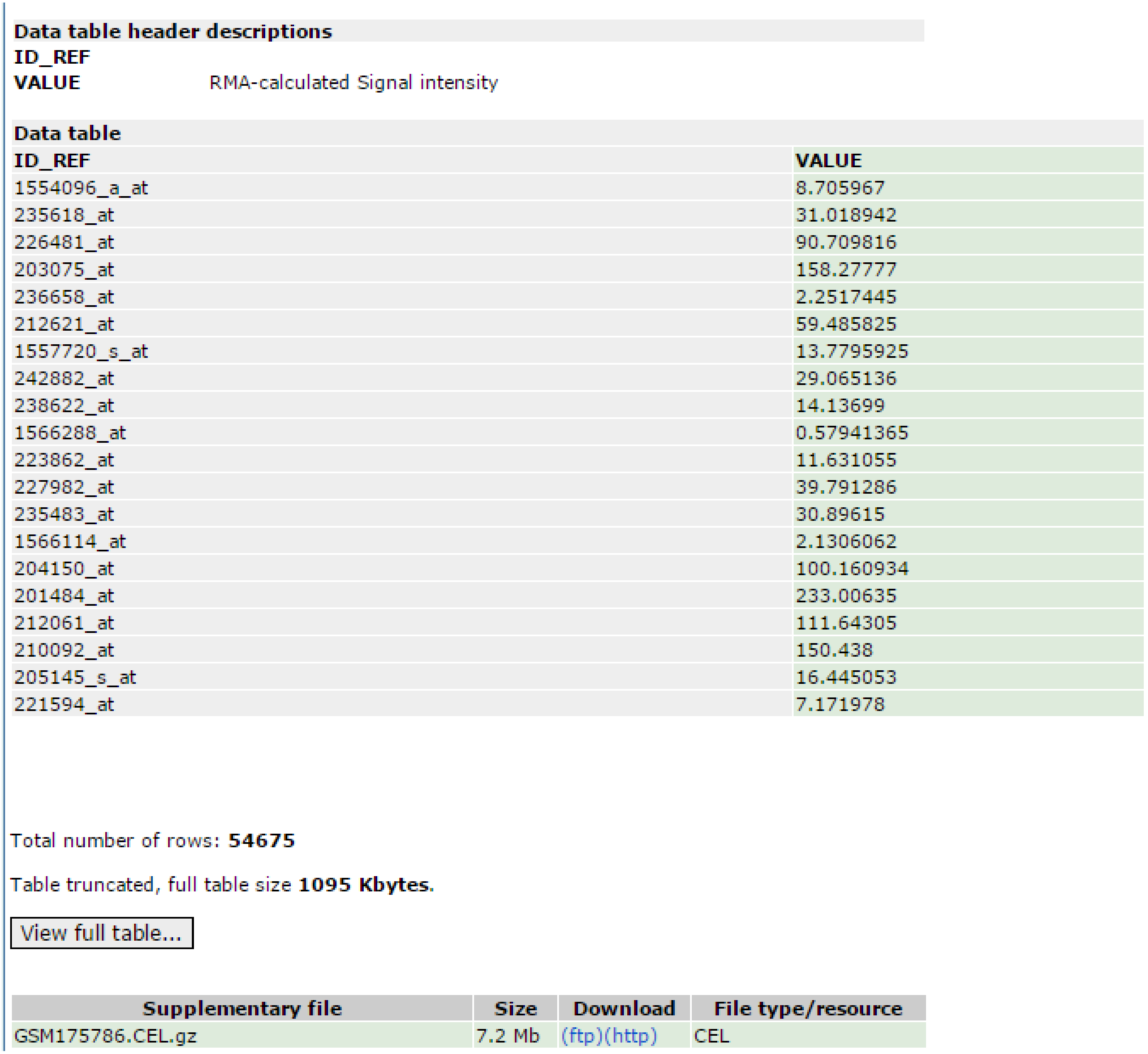
NCBI > GEO > **Accession Display** Not logged in | Login

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSM175786 GO

Sample GSM175786 Query DataSets for GSM175786

Status	Public on Apr 09, 2007
Title	Endometrium/Ovary 1 Disease
Sample type	RNA
Source name	Male/Female; Normal/Diseased
Organism	Homo sapiens
Characteristics	Tissue/Cell Line [C]: endometrium/ovary Disease/Normal or Treatment [C]: disease Gender: F Disease type: Endometriosis
Treatment protocol	Not treated
Growth protocol	None
Extracted molecule	total RNA
Extraction protocol	Trizol extraction of total RNA was performed according to the manufacturer's instructions.
Label	biotin
Label protocol	Biotinylated cRNA were prepared according to the standard Affymetrix protocol (Expression Analysis Technical Manual, 2001, Affymetrix).
Hybridization protocol	Following fragmentation, cRNA were hybridized for 16 hr at 45C on the GeneChip Human U133 plus 2.0 array. GeneChips were washed and stained in the Affymetrix Fluidics Station 450.
Scan protocol	GeneChips were scanned using the Affymetrix GeneChip Scanner 3000
Description	Disease
Data processing	The data were analyzed with robust multi-array (RMA) for background correction, normalization and polishing.
Submission date	Mar 19, 2007
Last update date	Jan 14, 2015
Contact name	Richard B Roth
E-mail	rroth@neurocrine.com
Organization name	Neurocrine Biosciences, Inc.
Department	Molecular Medicine
Street address	12790 El Camino Real
City	San Diego
State/province	CA
ZIP/Postal code	92130
Country	USA



Data table header descriptions

ID_REF	
VALUE	RMA-calculated Signal intensity

Data table

ID_REF	VALUE
1554096_a_at	8.705967
235618_at	31.018942
226481_at	90.709816
203075_at	158.27777
236658_at	2.2517445
212621_at	59.485825
1557720_s_at	13.7795925
242882_at	29.065136
238622_at	14.13699
1566288_at	0.57941365
223862_at	11.631055
227982_at	39.791286
235483_at	30.89615
1566114_at	2.1306062
204150_at	100.160934
201484_at	233.00635
212061_at	111.64305
210092_at	150.438
205145_s_at	16.445053
221594_at	7.171978

Total number of rows: **54675**

Table truncated, full table size **1095 Kbytes**.

Supplementary file	Size	Download	File type/resource
GSM175786.CEL.gz	7.2 Mb	(ftp)(http)	CEL