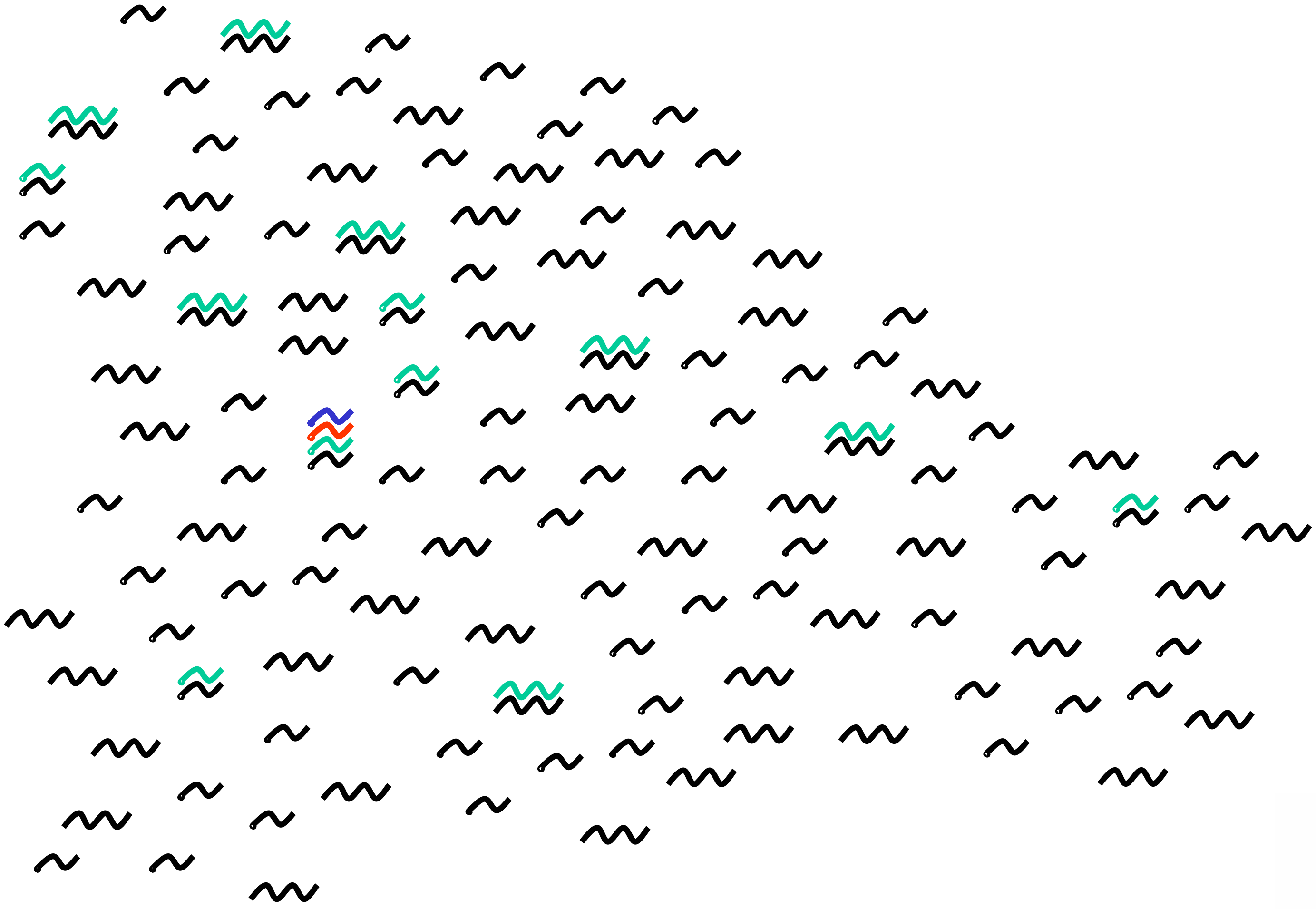
The EMBL Nucleotide Sequence Database:

Exploiting commonalities between records

Funded by:

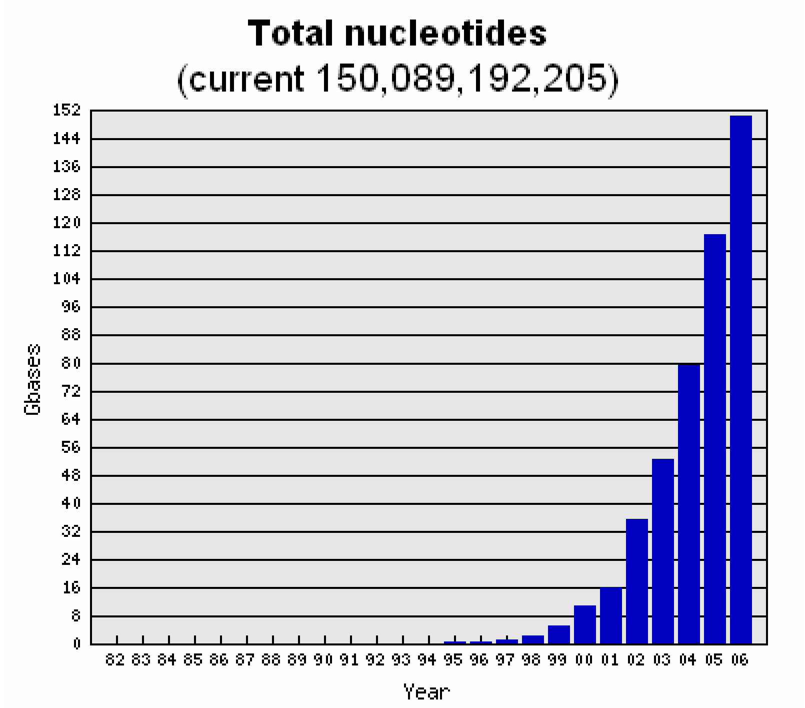


INSDC aims to gather and make freely available nucleotide sequence and annotation with comprehensive global coverage.

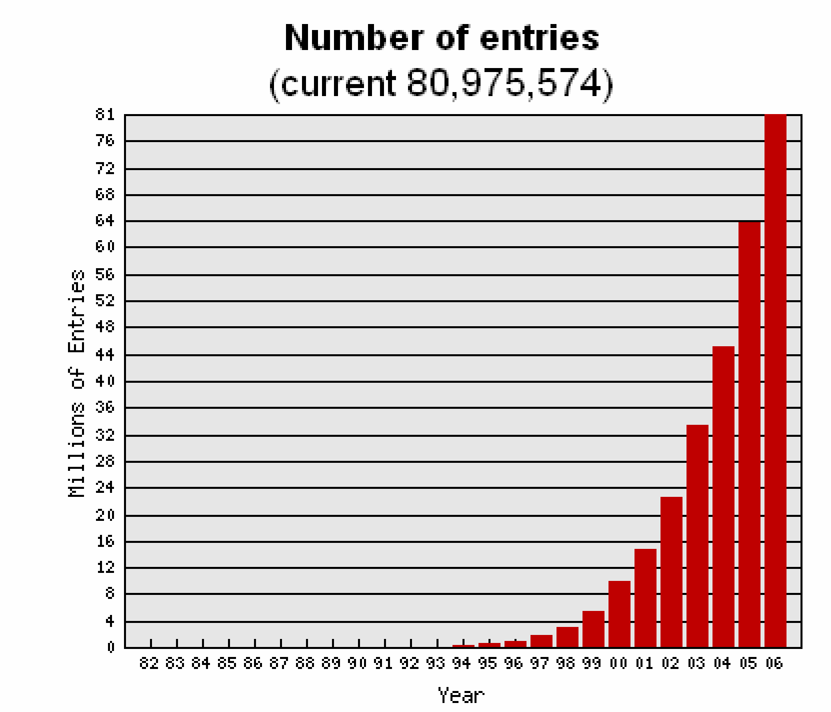
Ownership, and hence editorial control, of biological content of entries remains with the original submitting group.

Funded by:

Current database status



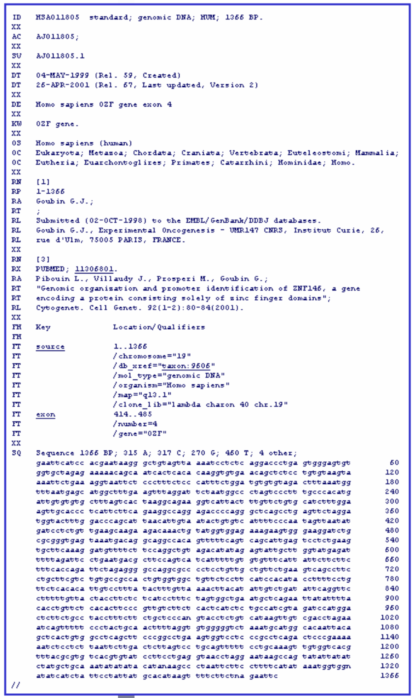
Total nucleotides
(current 150,089,192,205)



Number of entries
(current 80,975,574)

Funded by:

EMBL entry



```
ID   HSA011805  standard; genomic DNA; HUM; 1366 BP.
XX
AC   AJ011805;
XX
SV   AJ011805.1
XX
DT   04-MAY-1999 (Rel. 59, Created)
DT   26-APR-2001 (Rel. 67, Last updated, Version 2)
XX
DE   Homo sapiens OZF gene exon 4
XX
KW   OZF gene.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-1366
RA   Goubin G.J.;
RT   ;
RL   Submitted (02-OCT-1998) to the EMBL/GenBank/DDBJ databases.
RL   Goubin G.J., Experimental Oncogenesis - UMR147 CNRS, Institut Curie, 26,
RL   rue d'Ulm, 75005 PARIS, FRANCE.
XX
RN   [3]
RX   PUBMED; 11306801.
RA   Pibouin L., Villaudy J., Prosperi M., Goubin G.;
RT   "Genomic organization and promoter identification of ZNF146, a gene
RT   encoding a protein consisting solely of zinc finger domains";
RL   Cytogenet. Cell Genet. 92(1-2):80-84(2001).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1366
FT                   /chromosome="19"
FT                   /db_xref="taxon:9606"
FT                   /mol_type="genomic DNA"
FT                   /organism="Homo sapiens"
FT                   /map="q13.1"
FT                   /clone_lib="lambda charon 40 chr.19"
FT   exon            414..485
FT                   /number=4
FT                   /gene="OZF"
XX
SQ   Sequence 1366 BP; 315 A; 317 C; 270 G; 460 T; 4 other;
     gaattcatcc acgaataagg gctgtagtta aaatcctctc aggaccctga gtgggagtgt        60
     ggtgctagag aaaaacagca atcactcaca caaggtgtga acagctctcc tgtgtaagta       120
     aaattctgaa aggtaattct ccctttctcc catttctgga tgtgtgtaga ctttaaatgg       180
     tttaatgagc atggctttga agtttaggat tctaatggcc ctagtccctt tgcccacatg       240
     attgtgtgtg ctttagtcac taaggcagaa ggtcattact ttgttctgtg catctttgga       300
     agttgcaccc tcattcttca gaaggccagg agaccccagg gctcagcctg agttctagga       360
     tggtactttg gacccagcat taacattgta atactgtgtc attttcccaa tagttaatat       420
     gatcctctgt tgaagcaaga agacaaactg tatggtggag aaagaagtgg gaaggatctg       480
     cgcgggtgag taaatgacag gcaggccaca gtttttcagt cagcattgag tcctctgaag       540
     tgcttcaaag gatgttttct tccaggctgt agacatatag agtattgctt ggtatgagat       600
     ttttagattc ctgaatgacg cttccagtca tcatttttgt gtgtttcatt attcttcttc       660
     tttcaccaga ttctagaggg gccaggcgcc cctcctgttg ctgttctgaa gtcagccttc       720
     ctgcttcgtc tgtgccgcca ctgtggtggc tgttctcctt catccacata ccttttcctg       780
     ttctcacaca ttgtccttta tactttgtta aaacttacat attgtctgat attcaggttc       840
     ctttttgtta ctaccttctc tcatcctttc tagtggctga atgctcagaa ttatatttta       900
     cacctgttct cacacttccc gttgtcttct cactcatctc tgccatcgta gatccatgga       960
     ctcttctgcc tacctttctt ctgctcccan gtacctctgt cataagttgt cgacctagaa      1020
     atcagttttt ccctactgca acttttaggt gtgggggtct aaatgcatgg cacaattaca      1080
     gctcactgtg gcctcagctt cccggcctga agtggtcctc ccgcctcaga ctcccgaaaa      1140
     aatctcctct taattcttga ctcttagtcc tgcagttttt cctgcaaagt tgtggtcacg      1200
     tttacgcgtg tcacgtgtat ccttcctgag gtaacctagg aataagccag tatattatat      1260
     ctatgctgca aatatatata catanaagcc ctaattcttc cttttcatat aaatggtggn      1320
     atatcatcta ttcctattat gcacataagt tttcttctna gaattc                    1366
//
```

Identifier and description

```
ID   HSA011805  standard; genomic DNA; HUM; 1366 BP.
XX
AC   AJ011805;
XX
SV   AJ011805.1
XX
DT   04-MAY-1999 (Rel. 59, Created)
DT   26-APR-2001 (Rel. 67, Last updated, Version 2)
XX
DE   Homo sapiens OZF gene exon 4
XX
KW
```

Submission reference

```
RN   [1]
RP   1-1366
RA   Goubin G.J.;
RT   ;
RL   Submitted (02-OCT-1998) to the EMBL/GenBank/DDBJ databases.
RL   Goubin G.J., Experimental Oncogenesis - UMR147 CNRS, Institut Curie, 26,
RL
```

Bibliographic reference

```
RN   [3]
RX   PUBMED; 11306801.
RA   Pibouin L., Villaudy J., Prospe
RT   "Genomic organization and promo
RT   encoding a protein consisting solely of zinc finger domains";
RL
```

Source molecule

```
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1366
FT                   /chromosome="19"
FT                   /db_xref="taxon:9606"
FT                   /mol_type="genomic DNA"
FT                   /organism="Homo sapiens"
FT                   /map="q13.1"
FT                   /clone_lib="lambda charon 40 chr.19"
```

Feature annotation

```
FT   exon            414..485
FT                   /number=4
FT                   /gene="OZF"
```

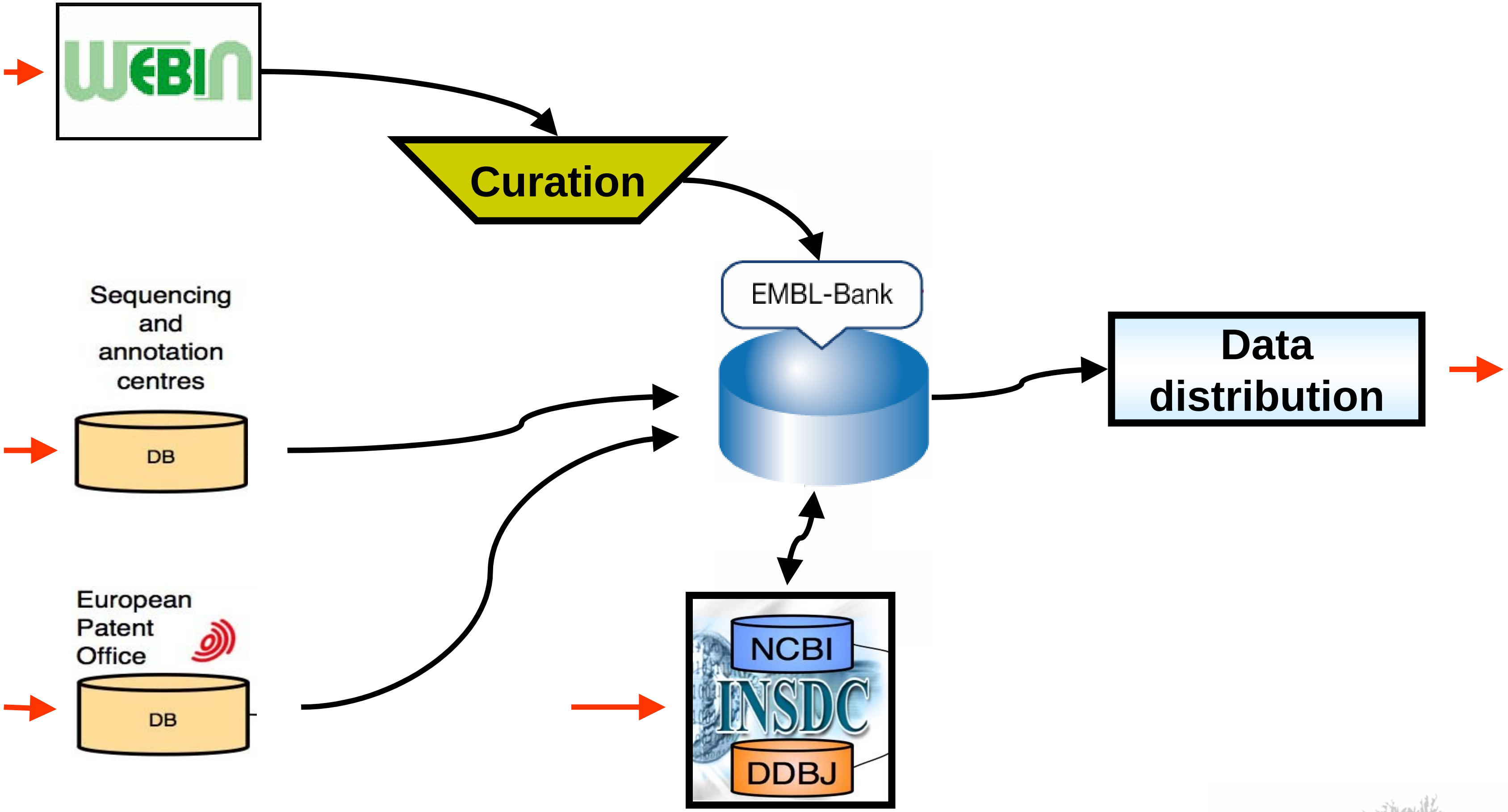
Sequence

```
SQ   Sequence 1366 BP; 315 A; 317 C; 270 G; 460 T; 4 other;
     gaattcatcc acgaataagg gctgtagtta aaatcctctc aggaccctga gtgggagtgt        60
     ggtgctagag aaaaacagca atcactcaca caaggtgtga acagctctcc tgtgtaagta       120
     aaattctgaa aggtaattct ccctttctcc catttctgga tgtgtgtaga ctttaaatgg       180
     tttaatgagc atggctttga agtttaggat tctaatggcc ctagtccctt tgcccacatg       240
     attgtgtgtg ctttagtcac taaggcagaa ggtcattact ttgttctgtg catctttgga       300
     agttgcaccc tcattcttca gaag                   agcctg agttctagga       360
     tggtactttg gacccagcat taac                   tcccaa tagttaatat       420
                                                  aagtgg gaaggatctg       480
                                  ggccaca gtttttcagt cagcattgag tcctctgaag       540
                                  aggctgt agacatatag agtattgctt ggtatgagat       600
```

Cross-reference

```
FT   source          1..1366
FT                   /db_xref="taxon:9606"
```

Data Flow



WEBIN

Curation

EMBL-Bank

Data distribution

Sequencing and annotation centres

DB

European Patent Office

DB

NCBI

INSDC

DDBJ

Funded by:

Data integration

- **49,323,034 entry-level cross-references**
- **12,787,002 feature-level cross-references**
- **further cross-references**
- **feature-level cross-references**

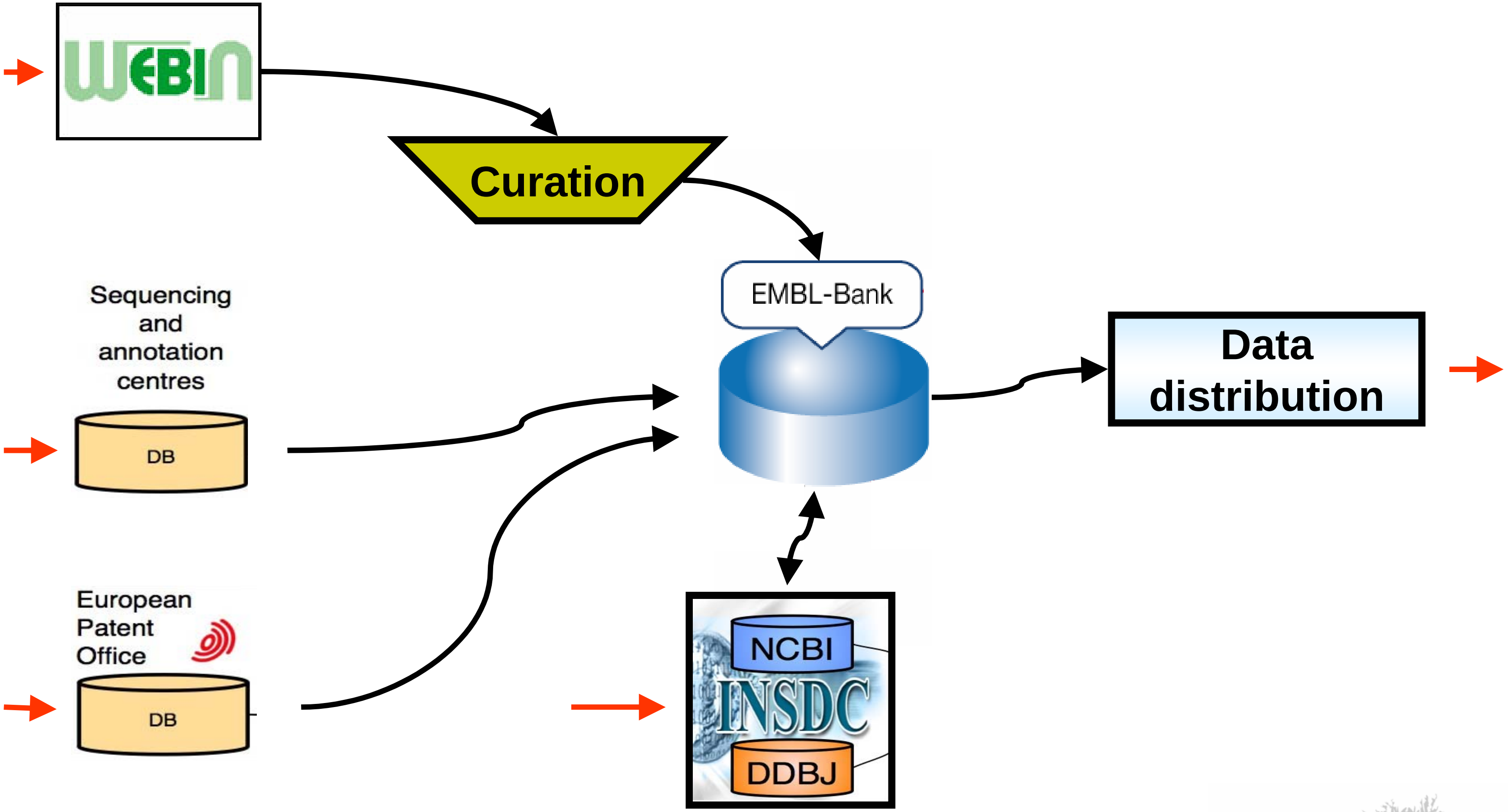
Funded by:

Data retrieval

- WWW
  - Sequence Retrieval System (SRS), srs.ebi.ac.uk
  - Simple sequence retrieval (Dbfetch), www.ebi.ac.uk/cgi-bin/emblfetch
  - Flatfile, INSDseq XML, EMBL XML, fasta, etc.
  - Whole genomes, www.ebi.ac.uk/genomes/
  - Sequence Version Archive, www.ebi.ac.uk/cgi bin/sva/sva.pl
- EBI sequence similarity search services
  - eg. http://www.ebi.ac.uk/Tools/homology.html
- FTP site
  - ftp.ebi.ac.uk/pub/databases/embl/
- E-mail file server, netserv@ebi.ac.uk
- Specialist data sets at users’ request (eg. EMBL CDS)

Funded by:

Data Flow



Funded by:

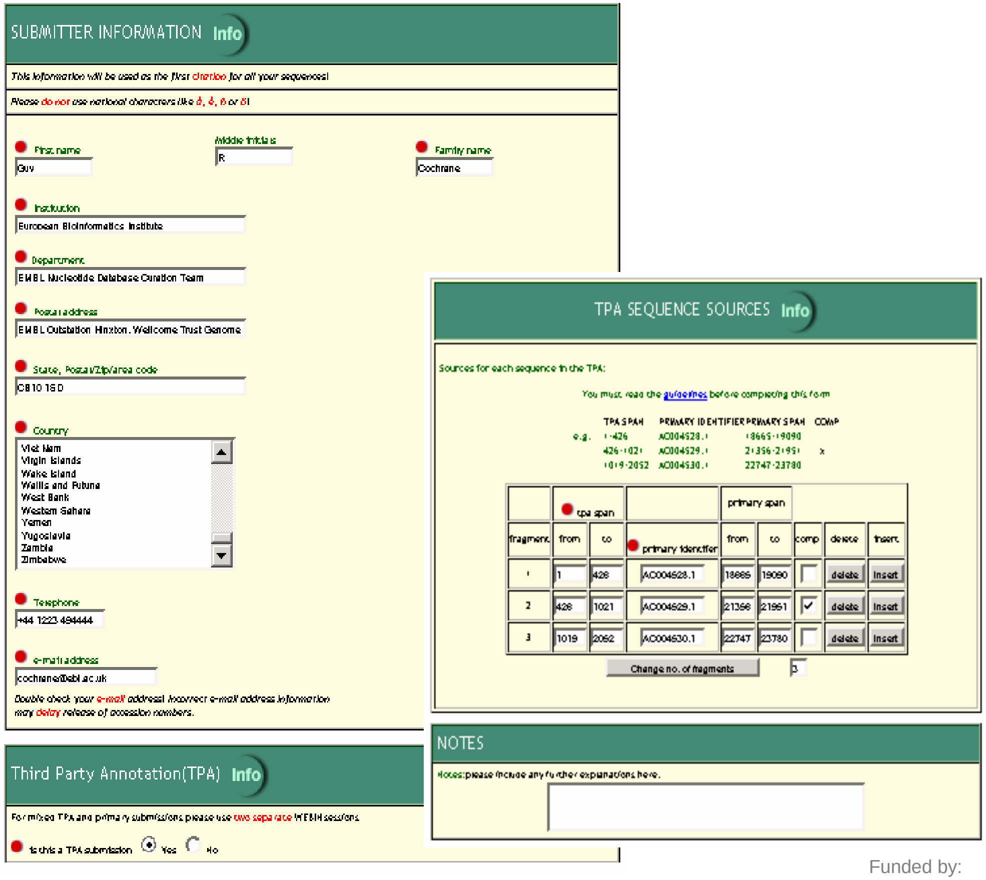
What is curation?

- ensuring compliance with annotation policies to maximise data consistency
- recommendation of appropriate nomenclatures
- maximising information content
- simplifying and accelerating submission procedure for submitters

Funded by:

Webin: Data submissions

- Submission of small numbers of entries
  - submitter moves through Web forms to submit each entry in turn, with some facility to copy from previous entries



SUBMITTER INFORMATION Info

This information will be used as the first citation for all your sequences!

Please do not use national characters like å, ä, ö or ß!

● First name: Guy
Middle initials: R
● Family name: Cochrane
● Institution: European Bioinformatics Institute
● Department: EMBL Nucleotide Database Curation Team
● Postal address: EMBL Outstation Hinxton, Wellcome Trust Genome
● State, Postal/Zip/area code: CB10 1SD
● Country:
Viet Nam
Virgin Islands
Wake Island
Wallis and Futuna
West Bank
Western Sahara
Yemen
Yugoslavia
Zambia
Zimbabwe
● Telephone: +44 1223 494444
● e-mail address: cochrane@ebi.ac.uk

Double check your e-mail address! Incorrect e-mail address information may delay release of accession numbers.

Third Party Annotation(TPA) Info

For mixed TPA and primary submissions please use two separate WEBIN sessions

● Is this a TPA submission ● Yes ○ No

TPA SEQUENCE SOURCES Info

Sources for each sequence in the TPA:

You must read the guidelines before completing this form

	TPA SPAN	PRIMARY IDENTIFIER	PRIMARY SPAN	COMP
e.g.	1-426	AC004528.1	18665-19090	
	426-1021	AC004529.1	21356-21951	x
	1019-2052	AC004530.1	22747-23780	

	● tpa span			primary span				
fragment	from	to	● primary identifier	from	to	comp	delete	insert
1	1	426	AC004528.1	18665	19090	☐	delete	Insert
2	426	1021	AC004529.1	21356	21951	☑	delete	Insert
3	1019	2052	AC004530.1	22747	23780	☐	delete	Insert

Change no. of fragments: 3

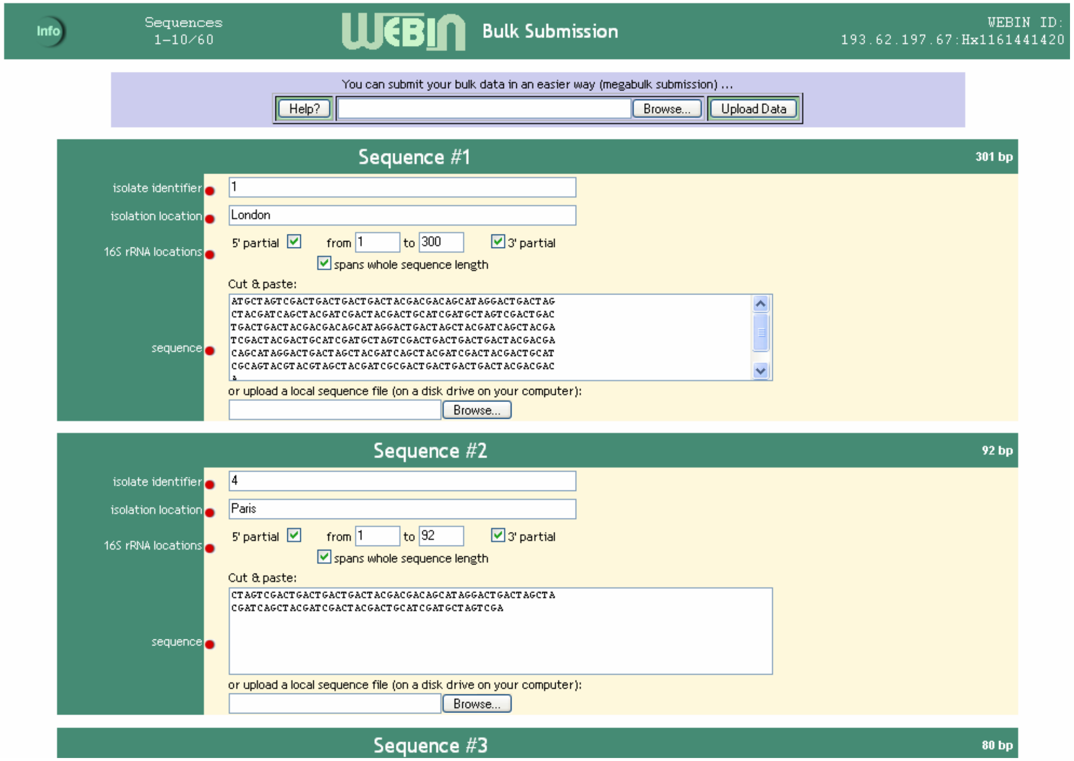
NOTES

Notes: please include any further explanations here.

Funded by:

Bulk submissions

- Submission of large numbers of entries with similar annotation
  - submission of representative sample entry
  - preparation of web form to recruit variable field data
  - upload of a file containing variable field information in a systematic format



You can submit your bulk data in an easier way (megabulk submission) ...

Sequence #1 301 bp

isolate identifier ● 1

isolation location ● London

16S rRNA locations ● 5' partial ☑ from 1 to 300 ☑ 3' partial

☑ spans whole sequence length

Cut & paste:

sequence ●

```
ATGCTAGTCGACTGACTGACTGACTACGACGACAGCATAGGACTGACTAG
CTACGATCAGCTACGATCGACTACGACTGCATCGATGCTAGTCGACTGAC
TGACTGACTACGACGACAGCATAGGACTGACTAGCTACGATCAGCTACGA
TCGACTACGACTGCATCGATGCTAGTCGACTGACTGACTGACTACGACGA
CAGCATAGGACTGACTAGCTACGATCAGCTACGATCGACTACGACTGCAT
CGCAGTACGTACGTAGCTACGATCGCGACTGACTGACTGACTACGACGAC
```

or upload a local sequence file (on a disk drive on your computer):

Sequence #2 92 bp

isolate identifier ● 4

isolation location ● Paris

16S rRNA locations ● 5' partial ☑ from 1 to 92 ☑ 3' partial

☑ spans whole sequence length

Cut & paste:

sequence ●

```
CTAGTCGACTGACTGACTGACTACGACGACAGCATAGGACTGACTAGCTA
CGATCAGCTACGATCGACTACGACTGCATCGATGCTAGTCGA
```

or upload a local sequence file (on a disk drive on your computer):

Sequence #3 80 bp

```
> a1_001 ; 28 ; 502 ; Beijing
atgctgatgcatgactcacgactagcactgactgacacgtaggacgacgacgactgacgatcgactgaca
ctgactgacatcgacgtacgacgatgcatcgatgcatcgatagacacatcacacagcacgtttatactac
acgtacgatgactgacgacgatcgatcggggactactacgactgactacagct
> a1_002 ; 12 ; 42 ; London
atgctgatgcatgactcacgactagcactgactgacacgtaggacgacgacgactgacgatcgactgac
actgactgacatcgacgtacgacgatgcatcgatgcatcgatagacacatcactttnnntttatactac
acgtacgatgactgacgacgatcgatcggggactactacgactgactacagct
> a1_003 ; 51 ; 91 ; Paris
atgctgatgcatgactcacgactagcactgactgacacgtaggacgacgacgactgacgatcgactgac
actgactgacatcgacgtacgacgatgcatcgatgcatcgatagacacatcacttttacgatatactac
acgtacgatgactgacgacgatcgatcggggactactacgactgactacagct
> a2_001 ; 80 ; 115 ; Tokyo
atgctgatgcatgactcacgactagcactgactgacacgtaggacgacgacgactgacgatcgactgac
actgactgacatcgacgtacgacgatgcatcgatgcatcgatagacacatcactttttttttatactac
acgtacgatgactgacgacgatcgatcggggactactacgactgactacagct
> b6_231 ; 92 ; 643 ; Shanghai
tactgactgacatcgacgtacgacgatgcatcgatgcatcgatagacacatcactttttttttatacta
atgtactgactgacatcgacgtacgacgatgcatcgatgcatcgatagacacatca
```

Funded by:

Curated submissions

Nucleic Acids Research, 2006, Vol. 34, Database issue

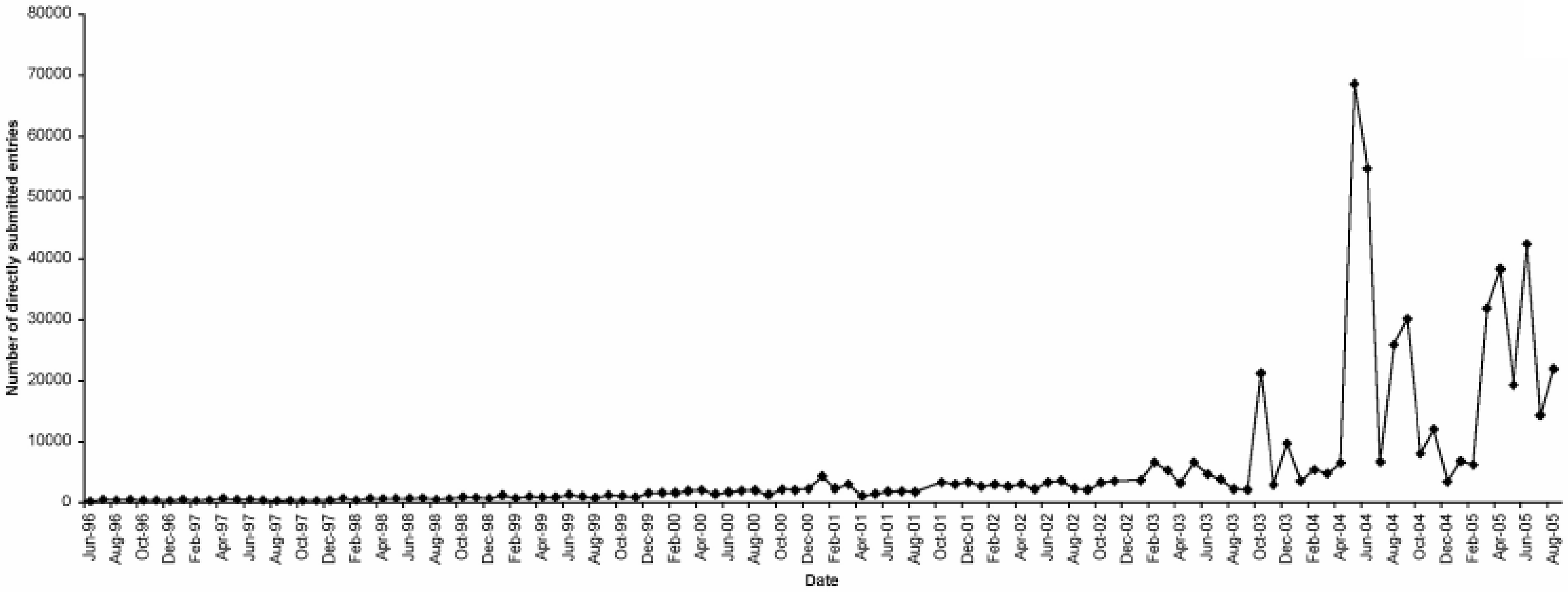
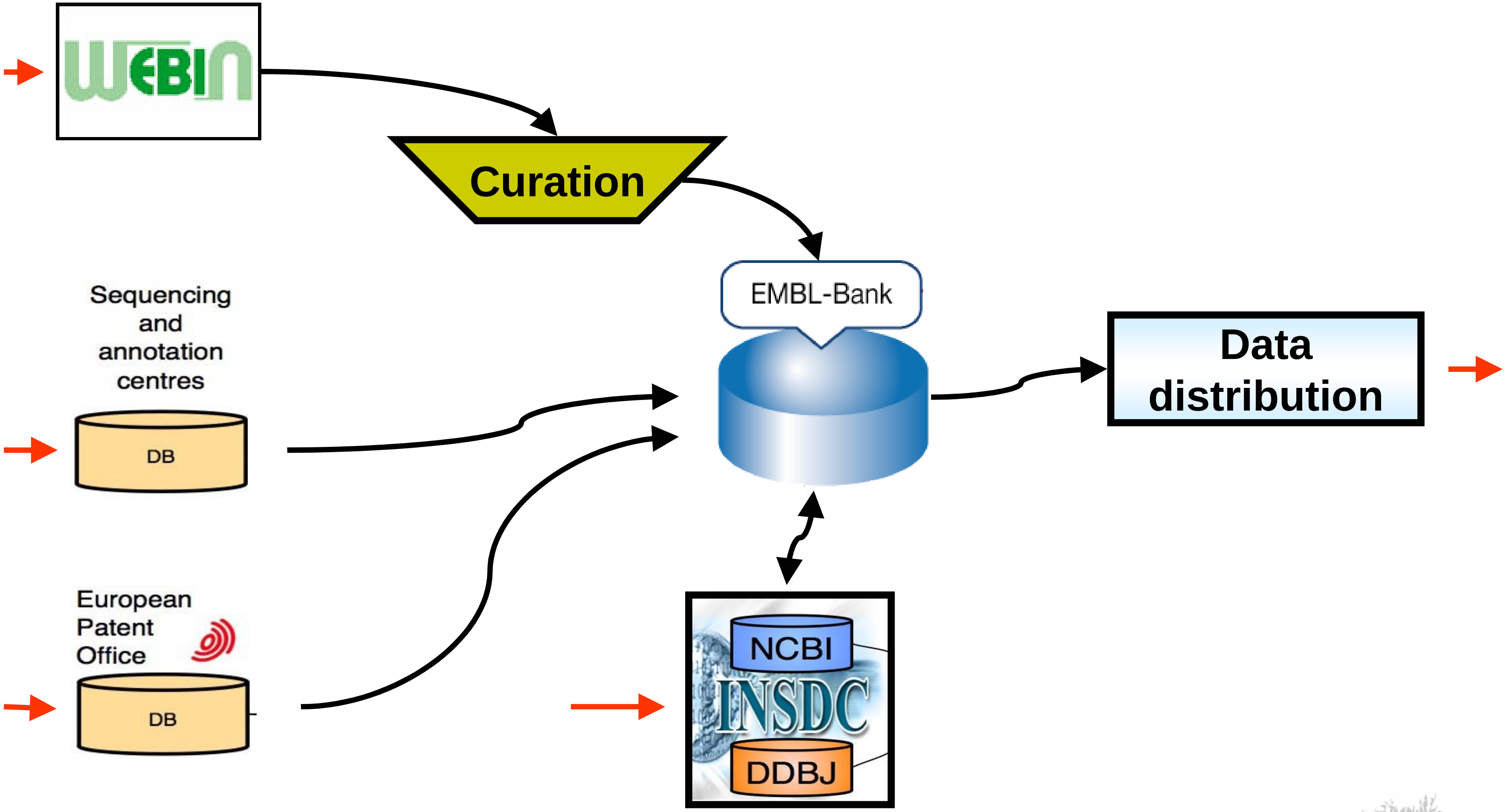


Figure 1. Increase in scale of direct submissions. Monthly newly submitted entry counts are shown.

Funded by:

Data Flow



WEBIN

Curation

Sequencing and annotation centres

DB

European Patent Office

DB

EMBL-Bank

Data distribution

NCBI

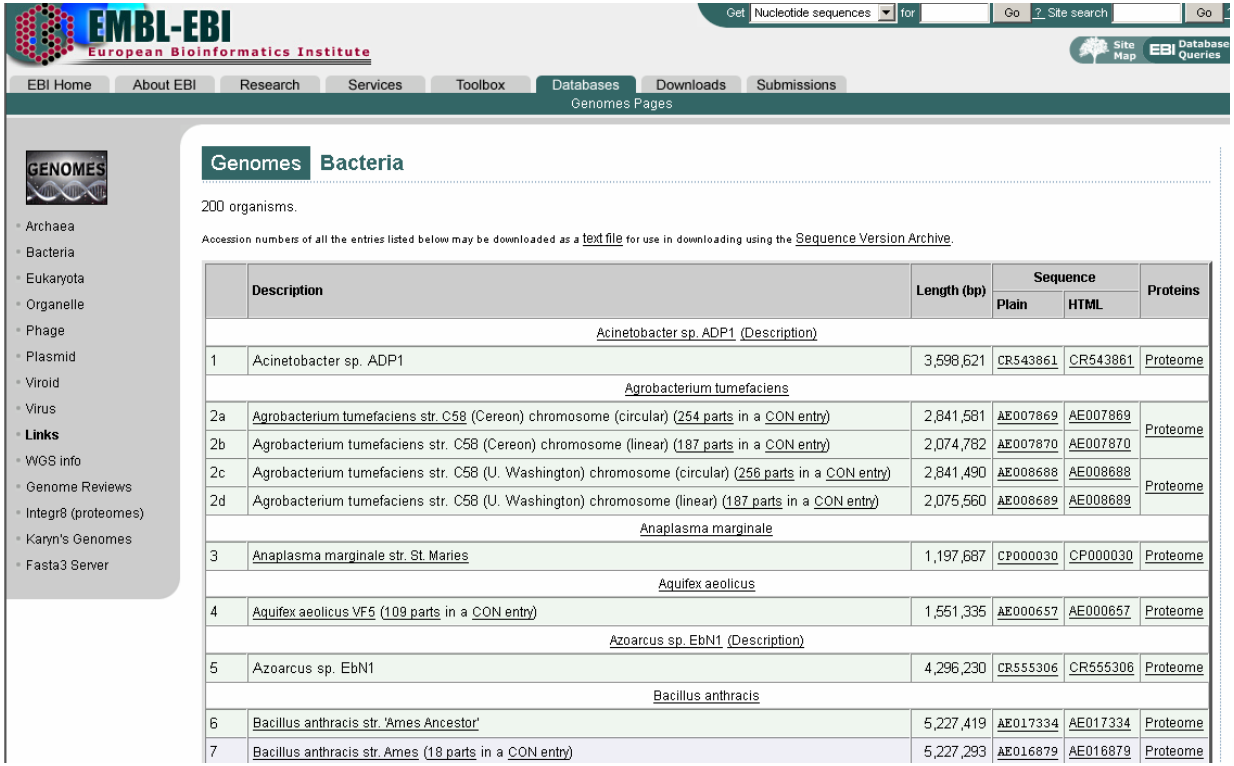
INSDC

DDBJ

Funded by:

Genomes

- Completely sequenced genomes and annotation
- 373 bacterial, 1212 viral, 50 eukaryotic, etc.



Genomes Bacteria

200 organisms.

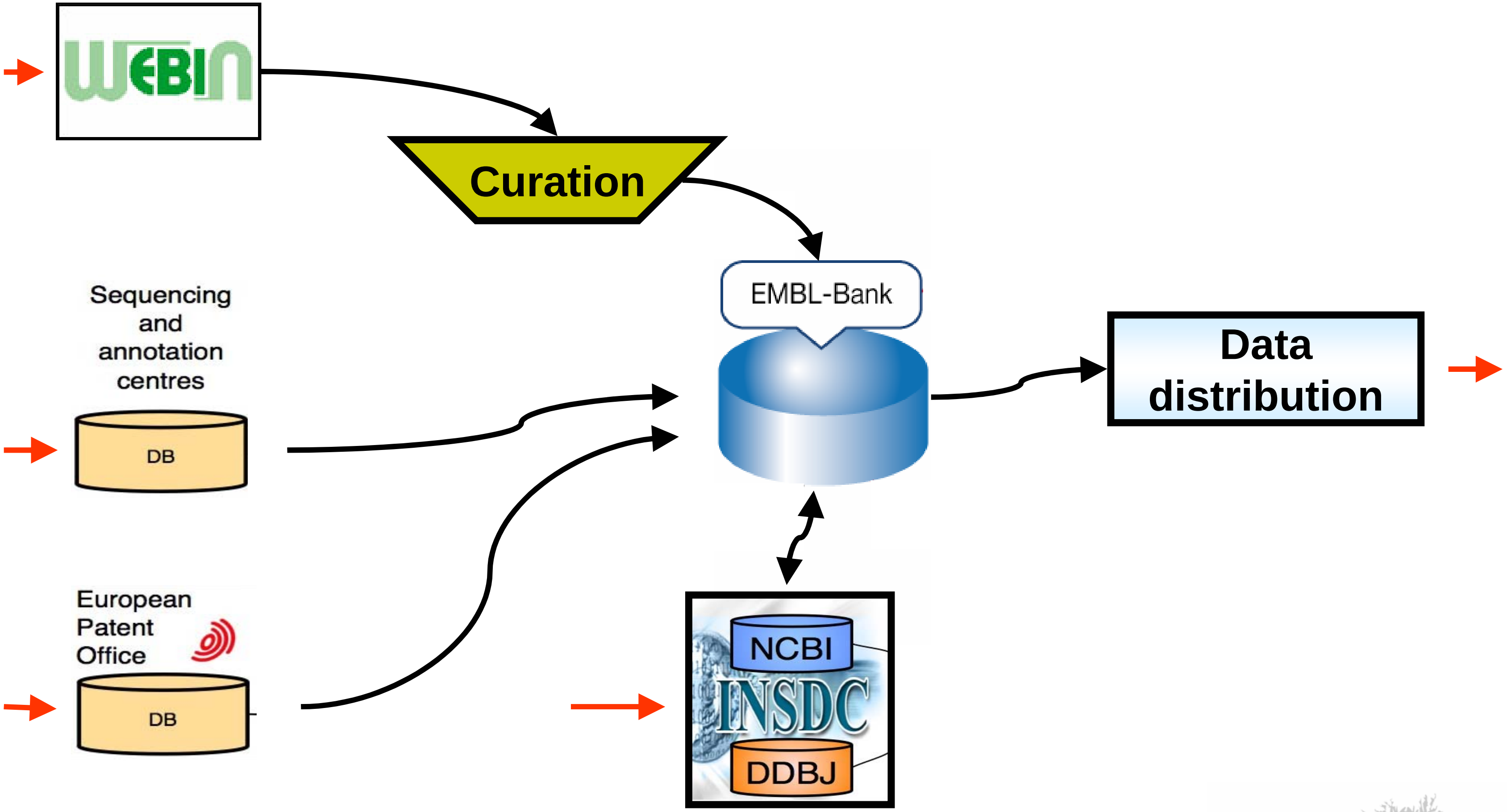
Accession numbers of all the entries listed below may be downloaded as a text file for use in downloading using the Sequence Version Archive.

	Description	Length (bp)	Sequence		Proteins
			Plain	HTML	
Acinetobacter sp. ADP1 (Description)					
1	Acinetobacter sp. ADP1	3,598,621	CR543861	CR543861	Proteome
Agrobacterium tumefaciens					
2a	Agrobacterium tumefaciens str. C58 (Cereon) chromosome (circular) (254 parts in a CON entry)	2,841,581	AE007869	AE007869	Proteome
2b	Agrobacterium tumefaciens str. C58 (Cereon) chromosome (linear) (187 parts in a CON entry)	2,074,782	AE007870	AE007870	
2c	Agrobacterium tumefaciens str. C58 (U. Washington) chromosome (circular) (256 parts in a CON entry)	2,841,490	AE008688	AE008688	Proteome
2d	Agrobacterium tumefaciens str. C58 (U. Washington) chromosome (linear) (187 parts in a CON entry)	2,075,560	AE008689	AE008689	
Anaplasma marginale					
3	Anaplasma marginale str. St. Maries	1,197,687	CP000030	CP000030	Proteome
Aquifex aeolicus					
4	Aquifex aeolicus VF5 (109 parts in a CON entry)	1,551,335	AE000657	AE000657	Proteome
Azoarcus sp. EbN1 (Description)					
5	Azoarcus sp. EbN1	4,296,230	CR555306	CR555306	Proteome
Bacillus anthracis					
6	Bacillus anthracis str. 'Ames Ancestor'	5,227,419	AE017334	AE017334	Proteome
7	Bacillus anthracis str. Ames (18 parts in a CON entry)	5,227,293	AE016879	AE016879	Proteome

- INSDC Project identifier to tie diverse entries into project
- Project metadata database

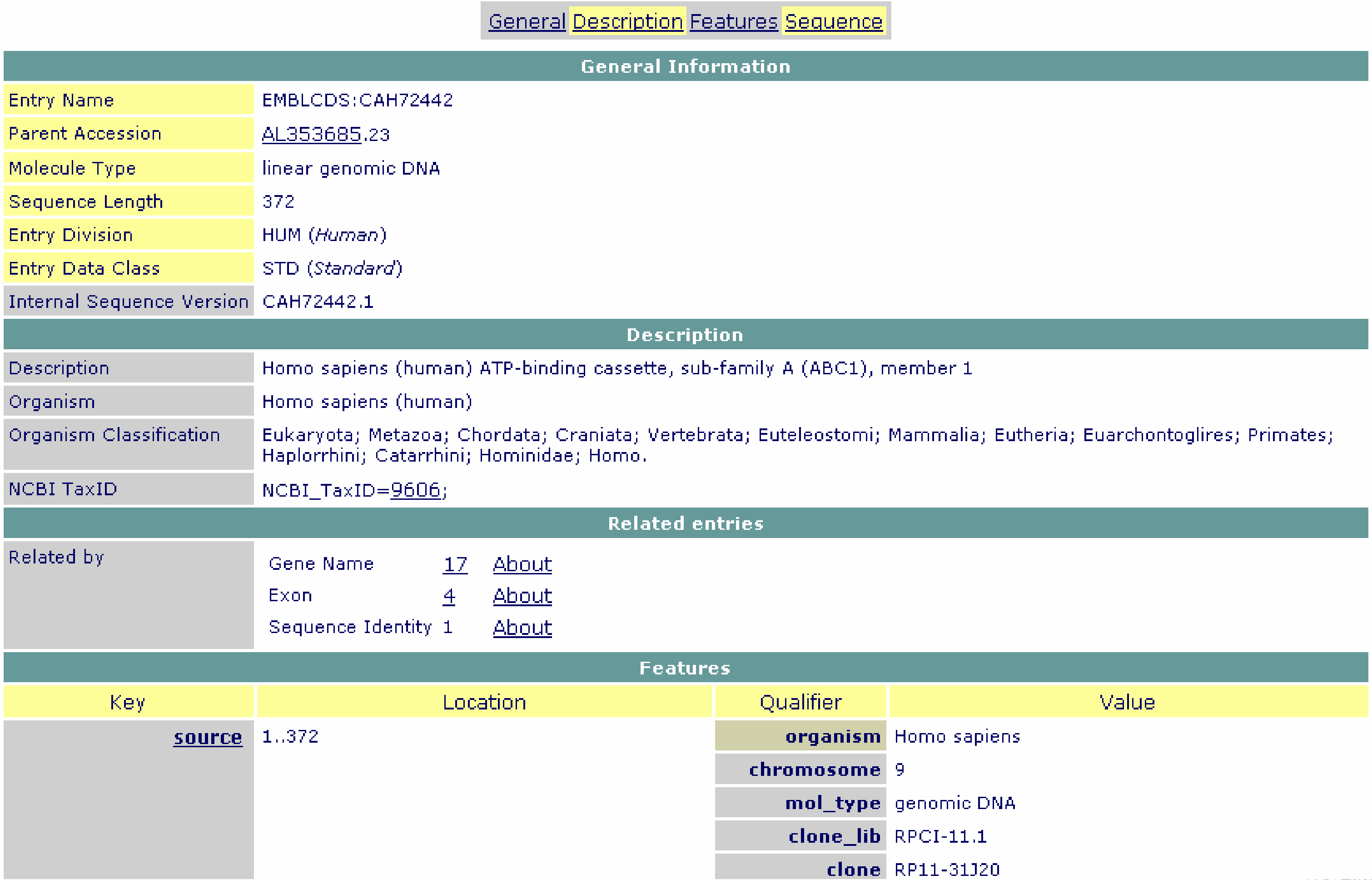
Funded by:

Data Flow



Funded by:

EMBL CDS groupings



General Information

Entry Name	EMBLCDS:CAH72442
Parent Accession	AL353685.23
Molecule Type	linear genomic DNA
Sequence Length	372
Entry Division	HUM (*Human*)
Entry Data Class	STD (*Standard*)
Internal Sequence Version	CAH72442.1

Description

Description	Homo sapiens (human) ATP-binding cassette, sub-family A (ABC1), member 1
Organism	Homo sapiens (human)
Organism Classification	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
NCBI TaxID	NCBI_TaxID=9606;

Related entries

Related by			
	Gene Name	17	About
	Exon	4	About
	Sequence Identity	1	About

Features

Key	Location	Qualifier	Value
source	1..372	**organism**	Homo sapiens
		chromosome	9
		mol_type	genomic DNA
		clone_lib	RPCI-11.1
		clone	RP11-31J20

Funded by:

EMBL CDS grouping

cds	join(complement(AL353685.23:101137..101258), complement(AL353685.23:99749..99867), complement(AL353685.23:78380..78510))

gene	ABCA1
locus_tag	RP11-413C10.7-004
standard_name	OTTHUMP00000021835
product	ATP-binding cassette, sub-family A (ABC1), member 1
db_xref	*GOA:Q5VYS1*
db_xref	UniProtKB/TrEMBL:Q5VYS1
protein_id	CAH72442.1
translation	>CAH72442.1 MPSAGTLPWVQGIICNANNPCFRYPTPGEAPGVVGNFNKSIVA SMKDMRKVLRTLQQIKKSSSNLKLQDFLVDNETFSGFLYHNLS KVS

Sequence

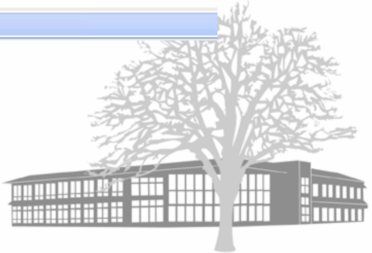
Characteristics	**Length:** 372 BP, **A Count:**96, **C Count:**83, **G Count:**96, **T Count:**97, **Others Count:**0 **CRC:**1405876788

Sequence

```
>CAH72442
atggatttgttaaagtttccaacaactccgggagcctccccaggagtcgggtaacggaaa
caggggttgttggcattacagataatcccctgaacccaaggaagtgttcctgcagagggc
atttgagctggatttcttgatctgctgtaatgttctcagaactttgcgcatgtccttcat
gctggtgtctttctggctgtataaaagaagcctccgagcatctgagaacaggcgagccac
atcagcttaccttgtggagaatgacatcagccctcagcatcttgtccacagtagactttg
ggagagagaggttgtgatacaggaacccagagaaggtttcattgtccaccaggaaatctt
gaagcttcaagt
```

General Description Features Sequence

Funded by:



People

- EMBL data submissions and curation
  - Karyn Duggan, Sheila Plaister, Bob Vaughan, Gaurab Mukherjee, Sumit Bhattacharyya, Ruth Akhtar, Kirsty Bates, Nadeem Faruque, Nicola Althorpe, Paul Browne, Philippe Aldebert, Ruth Eberhardt, Guy Cochrane
- EMBL database programmers
  - Carola Kanz, Dan Wu, Charles Lee, Dariusz Lorenc, Francesco Nardone, Rasko Leinonen, Alastair Baldwin, Quan Lin, Lawrence Bower, Siamak Sobhany, Matias Castro, Weimin Zhu
- Genome Reviews
  - Peter Sterk, Paul Kersey
- Database development and coordination
  - Tamara Kulikova, Guy Cochrane, Carola Kanz, Weimin Zhu, Rolf Apweiler
- External services team
- DDBJ and GenBank
- Cross-referring databases
- Submitters