

WS2017/2018

F1-Praktikum

„Genomforschung und Sequenzanalyse - Einführung in Methoden der Bioinformatik-“

PRINT ISSN: 0305-1048
ONLINE ISSN: 1362-4962

Nucleic Acids Research

VOLUME 45 DATABASE ISSUE JANUARY 4 2017
<https://academic.oup.com/nar>



OXFORD
UNIVERSITY PRESS

Open Access
No barriers to access – all articles freely available online



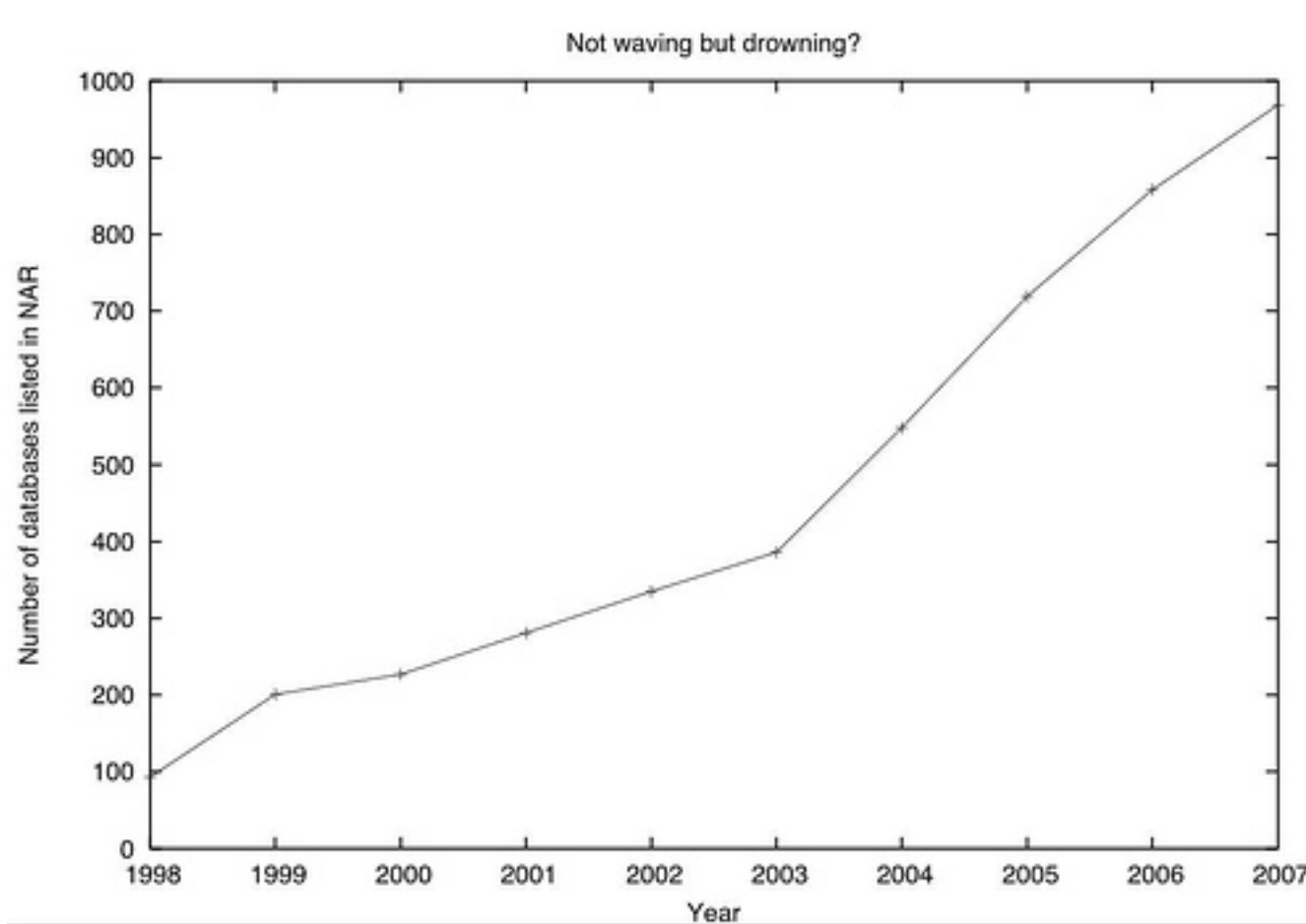
Datenbanken in der Molekularbiologie

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb,
 ARR, AsDb, BBDB, BCGD, Beanref, Biolmage,
 BioMagResBank, BIOMDB, **BLOCKS**, BovGBASE,
 BOVMAP, BSORF, BTKbase, CANSITE, CarbBank,
 CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP,
 ChickGBASE, Colibri, COPE, **CottonDB**, CSNDB, CUTG,
 CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb,
 Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC,
 ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db,
ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView,
 GCRDB, GDB, GENATLAS, Genbank, GeneCards,
 Geline, GenLink, GENOTK, GenProtEC, GIFTS,
 GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB,
 HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD,
 HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB,
 HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat,
 KDNA, **KEGG**, Klotho, LGIC, MAD, MaizeDb, MDB,
 Medline, Mendel, MEROPS, **MGDB**, MGI, MHCPEP5
 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us,
 MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase,
 OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB,
 PDD, **Pfam**, PhosphoBase, PigBASE, PIR, PKR, PMD,
 PPDB, PRESAGE, PRINTS, **ProDom**, Prolysis, PROSITE,
 PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE,
 SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase,
 etc... !!!!

Datenbanken in der Molekularbiologie

- >> 1000 unterschiedliche DBs
- normalerweise im Web
- unterschiedliche Größe: < 100 kb bis > 10 Gb
 - DNA > 10 Gb
 - Protein 1 Gb
 - 3D Struktur 5 Gb
- Update: täglich bis jährlich
- DB-Typen:
 - primär (Genbank, EMBL...)
 - abgeleitet (InterPro, PFAM....)
 - organismenspezifisch (Hefe, Arabidopsis ...)
 - datenspezifisch (Struktur, Expression, Pathways ...)

Datenbanken in der Molekularbiologie





Sequenz-Datenbanken

- komplette Übersicht: Januar-Ausgabe von Nucleic Acids Research

ISSN 0305-182X

Nucleic Acids Research

VOLUME 45, SUPPLEMENT 1, JANUARY 4, 2017
<http://nar.oupjournals.org>



z. B.

„Genbank“

„Flybase“

„Wanda“: a library of duplicated fish genes“

„ENZYME“

- <http://nar.oupjournals.org>

Datenbanken in der Molekularbiologie

- die wichtigsten Kategorien:

Literatur

Sequenzen

Genome

Proteinfamilien

Mutationen/Polymorphismen

3D Strukturen

Proteomics/2D-Gel, MS

Transcriptomics /Microarrays

Metabolische Netzwerke

Regulatorische Netzwerke

A A + OUP www.oxfordjournals.org/nar/database/c/

YouTube google Outlook Web App JGU NCBI Unimail ilias 78 molgen UCSC LE

Inbox - Outlook

OXFORD
ACADEMIC | Journals

You are here: [NAR Journal Home](#) » Database Summary Paper Categories

NAR Database Summary Paper Category List

- [Nucleotide Sequence Databases](#)
- [RNA sequence databases](#)
- [Protein sequence databases](#)
- [Structure Databases](#)
- [Genomics Databases \(non-vertebrate\)](#)
- [Metabolic and Signaling Pathways](#)
- [Human and other Vertebrate Genomes](#)
- [Human Genes and Diseases](#)
- [Microarray Data and other Gene Expression Databases](#)
- [Proteomics Resources](#)
- [Other Molecular Biology Databases](#)
- [Organelle databases](#)
- [Plant databases](#)
- [Immunological databases](#)
- [Cell biology](#)

...ab hier weiter suchen

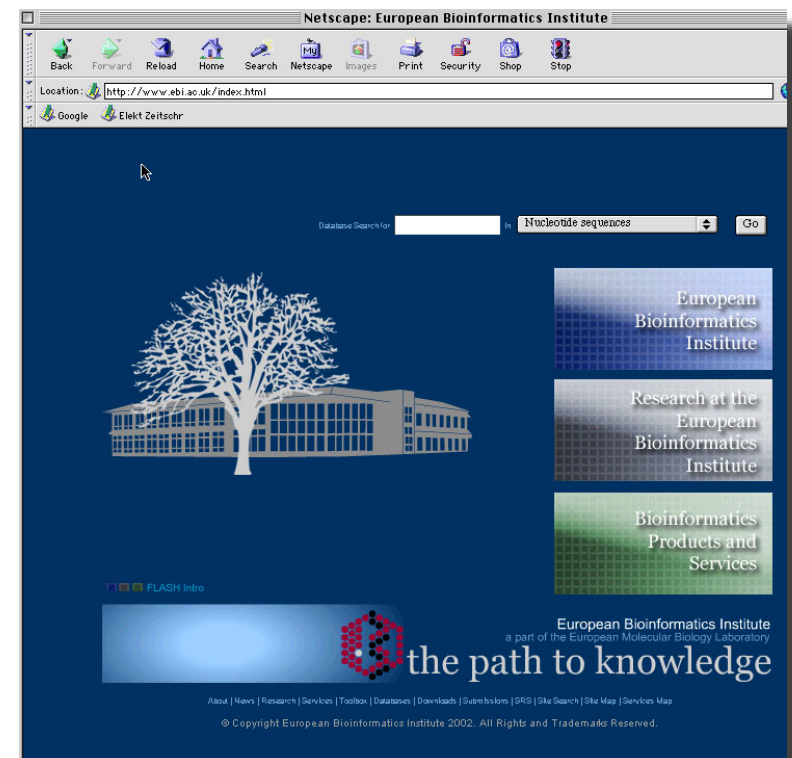
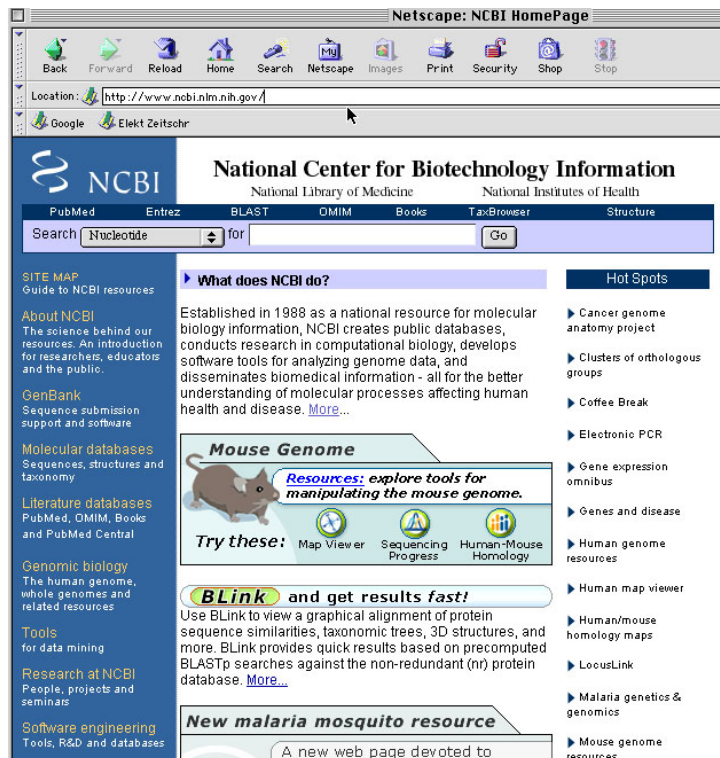
Datenbanken in der Molekularbiologie

<http://www.ncbi.nlm.nih.gov/>

National Center for Biotechnology Information,
Am NIH, Bethesda, Maryland, USA

<http://www.ebi.ac.uk>

European Bioinformatics Institute,
Sanger Campus, Hinxton, GB



Integrierte Such-Werkzeuge!

neuroglobin – GQuery: Global Cross-database NCBI search – NCBI

https://www.ncbi.nlm.nih.gov/gquery/?term=neuroglobin

Google YouTube Outlook Web App JGU NCBI Unimail ilias 78 molgen UCSC LEO News El-Zeitschr

NCBI Resources How To Sign in to NCBI

Search NCBI databases [Help](#)

neuroglobin Search

Results found in 22 databases for "neuroglobin"

Literature			Genes		
Books	5	books and reports	EST	3	expressed sequence tag sequences
MeSH	1	ontology used for PubMed indexing	Gene	383	collected information about gene loci
NLM Catalog	2	books, journals and more in the NLM Collections	GEO DataSets	0	functional genomics studies
PubMed	521	scientific & medical abstracts/citations	GEO Profiles	4,750	gene expression and molecular abundance profiles
PubMed Central	737	full-text journal articles	HomoloGene	1	homologous gene sets for selected organisms
			PopSet	1	sequence sets from phylogenetic and population studies
			UniGene	20	clusters of expressed transcripts
Health			Proteins		
ClinVar	10	human variations of clinical significance	Conserved Domains	0	conserved protein domains
dbGaP	0	genotype/phenotype interaction studies	Protein	627	protein sequences
GTR	1	genetic testing registry	Protein Clusters	0	sequence similarity-based protein clusters
MedGen	0	medical genetics literature and links	Structure	45	experimentally-determined biomolecular structures
OMIM	3	online mendelian inheritance in man			
PubMed Health	0	clinical effectiveness, disease and drug reports	Chemicals		
Genomes			BioSystems	115	molecular pathways with links to genes, proteins and chemicals
Assembly	0	genome assembly information	PubChem BioAssay	0	bioactivity screening studies
BioProject	0	biological projects providing data to NCBI			chemical information with structures, information and
BioSample	0	descriptions of biological source materials			
Clone	931	genomic and cDNA clones			

www.ncbi.nlm.nih.gov/

Der Einstieg in die Suche...

watson jd

Myoglobin

j mol evol

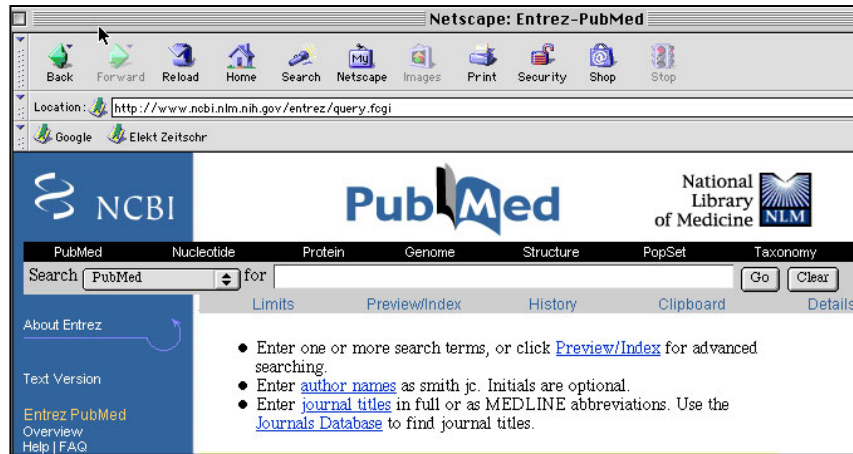
3M syndrome

P02998

mouse gtpase

Exkurs....

Literatur-Suche



- PubMed =
Public Medline /NCBI

- Suchdienst der Natl. Library of Medicine
- Medizin/ Life Sciences (aber nicht z. B. „Spezielle Botanik“)
- ca 19 000 Journals
- > 27 Mio. Einträge, Zeitschriften bis 1946 zurück
- Verbindung zu online-Zeitschriften > Download!!
- täglicher update
- Suchbegriffe einfach eingeben
(Boole'sche Verknüpfungen: „AND“, „OR“, „NOT“; Truncation: „*“)

Literatur-Suche

The screenshot shows the ISI Web of Knowledge portal. At the top is a green header with the text "ISI Web of KnowledgeSM" and a navigation bar with "Products & Features" and a "GO" button. Below the header, there is a green circular icon with a checkmark and the text "Take the next step with ISI Web of Knowledge". Below this, there is a section titled "Information when and how you want it." with links for "here" and "here" to view recorded training. Further down, there are links for "More information", "Notices", "Help", and "Tutorial". The main content area is divided into three sections: "CrossSearch" with a search bar containing "neuroglobin" and a "SEARCH" button; "Searchable Database Products" with a "GO" button; and "Analytical Tools" with a "GO" button. The "Searchable Database Products" section lists "Web of Science" with sub-items: "Science Citation Index Expanded", "Social Sciences Citation Index", and "Arts & Humanities Citation Index". The "Analytical Tools" section lists "Journal Citation Reports" with the sub-item "Journal performance metrics, including Impact Factor". The "Other Resources" section lists "ISI HighlyCited.com" with the sub-item "Author biographies and bibliographies".

- ISI Web of Science

<http://portal.isiknowledge.com/portal.cgi>

- **Journal Citation Index!! > Impact Factor**
- auch andere DB als Medline: BIOSIS previews etc etc
- Definition des Suchzeitraums
- Analytische Werkzeuge
- nette Spielereien: Biographien der „highly cited personalities“

Der Journal Impact-Factor (JIF)

- soll messen, wie oft andere Zeitschriften die Artikel der betrachteten Zeitschrift zitiert haben
- Ansehen der Fachzeitschrift > Qualität der Arbeit u. des Autors!!

Berechnung [\[Bearbeiten\]](#)

aus Wikipedia

Die Berechnung des *Journal Impact Factors* (JIF) erfolgt innerhalb einer Zwei-Jahres-Spanne^[6] nach folgender Formel:

$$\frac{\text{Zahl der Zitate im Bezugsjahr auf die Artikel der vergangenen zwei Jahre}}{\text{Zahl der Artikel in den vergangenen zwei Jahren}}$$

Daraus folgt: Es kann keinen solchen Impact Factor für ein noch nicht abgelaufenes Jahr geben. Beispiel: Eine Zeitschrift hat in den Jahren 2006–07 insgesamt 116 Artikel publiziert (A), im Jahr 2008 wurden Artikel aus dieser Zeitschrift insgesamt 224 mal zitiert (B), daraus ergibt sich für 2008 ein Impact Factor der Zeitschrift von 1,931 (B/A).

- vielfältig kritikwürdig!

Literatur-Suche



- grösster Vorteil: zeigt Link zu allen, die einen bestimmten Artikel zitiert haben! („Was haben die zu meinen Daten gesagt?“)

Literatur-Suche

There are currently no queries defined - please use the form below for setting them up.

New PubMed query:

Alias: (replace with **descriptive** term)

search term	search field	connector
<input type="text" value="neuroglobin"/>	<input type="text"/>	AND
<input type="text" value="cytoglobin"/>	<input type="text"/>	AND
<input type="text" value="globin AND developme"/>	<input type="text"/>	AND
<input type="text" value="development AND hypc"/>	<input type="text"/>	AND
<input type="text"/>	<input type="text"/>	

Type: ☒ PubMed ☐ PubMed Neighbour
☐ Nucleotide ☐ Nucleotide Neighbour

Terms:



- durchsucht PubMed und ENTREZ-DBs > Alarm!!

Literatur-Suche

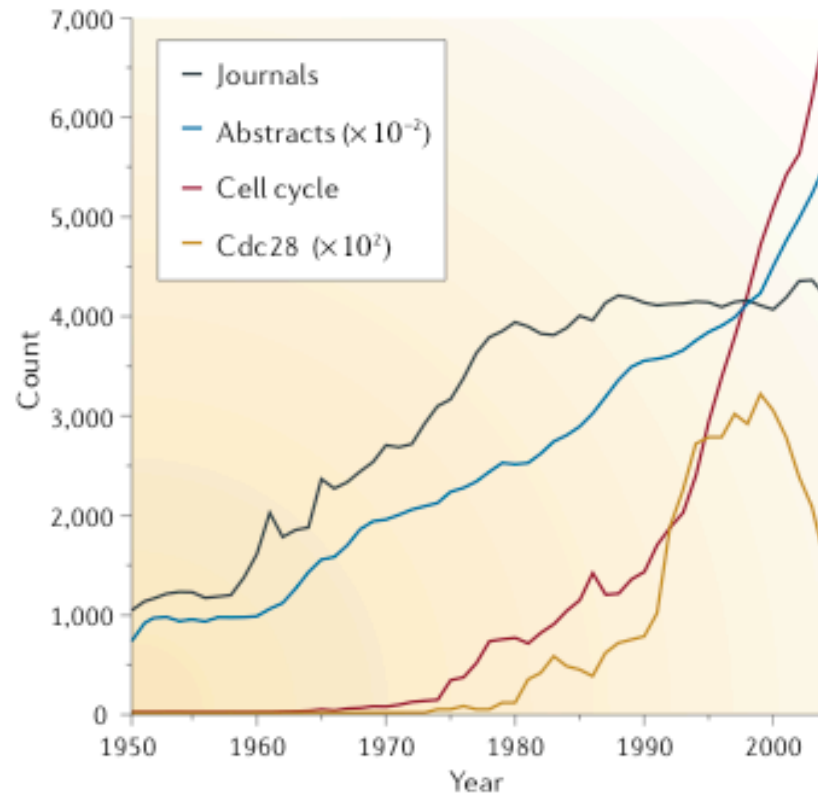


Figure 1 | **Growth of Medline.** The numbers of journals, papers (as represented by Medline abstracts), papers on the cell cycle and papers on Cdc28 that were published each year from 1950 to 2005 are shown. An average for 3 years was calculated for the Cdc28 curve because of much lower numbers. The number of new papers that were published each year continues to increase, especially on certain topics such as the cell cycle, for which it is no longer possible to read all new papers that are published. By contrast, specific proteins that are 'hot' at one point in time tend to lose their popularity later, as exemplified by Cdc28.

Zunehmend
ein Problem...

Lit-Suche > Discovery

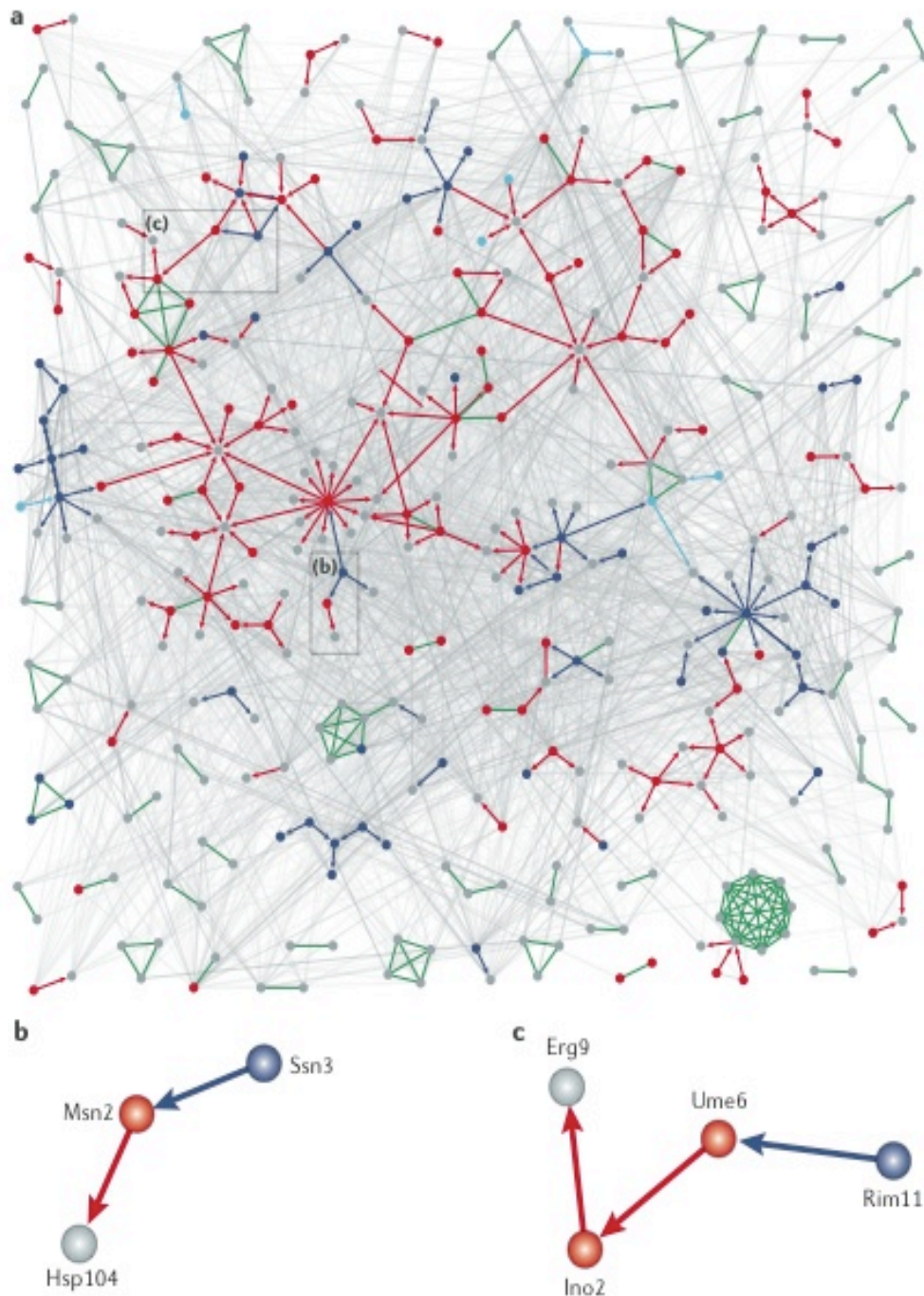
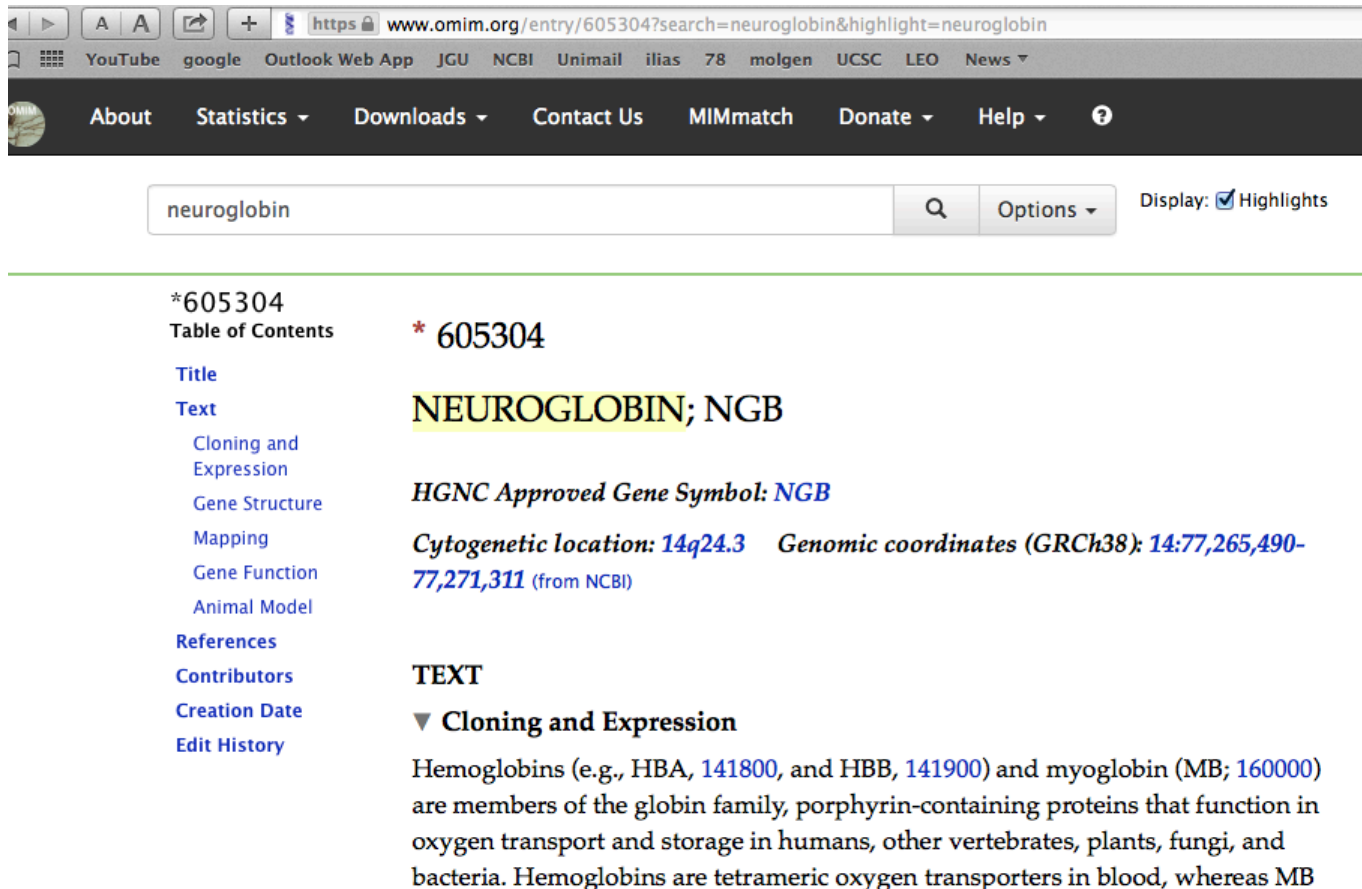


Figure 3 | **A literature-derived network for yeast.** **a** | A yeast protein network was derived that applied information-extraction approaches to all abstracts that are stored in Medline, using both a statistical co-occurrence method⁵⁴ and a natural-language-processing (NLP)-based one⁶². Functional associations that were derived from co-occurrence are shown in shades of grey according to the level of confidence that was achieved. The NLP method extracts four types of relationship: stable physical interactions (green), regulation of expression (red), phosphorylation (dark blue) and dephosphorylation (light blue). The proteins (circles) are coloured according to their functional annotation: (co-)regulators of expression (red), kinases and cyclins (dark blue), phosphatases (light blue) and other proteins (grey). A version of this figure that includes all protein names is available in the [supplementary information S1](#) (figure). **b,c** | Examples of unpublished relationships that can be inferred from the network. From the network we can infer that Ssn3 probably influences Hsp104 expression through phosphorylation of Msn2 (**b**). In addition, Ume6 probably regulates Erg9 expression and Rim11 is predicted to regulate the expression of both Ino2 and Erg9 (**c**). None of these hypotheses has been tested experimentally.

OMIM: eine spezielle Literatur-Datenbank

Online Mendelian Inheritance of Man
= Katalog menschlicher Gene und ihrer Erkrankungen



The screenshot shows a web browser window with the URL <https://www.omim.org/entry/605304?search=neuroglobin&highlight=neuroglobin>. The browser's address bar and search bar both contain the word "neuroglobin". The OMIM website header includes navigation links: About, Statistics, Downloads, Contact Us, MIMmatch, Donate, and Help. Below the header, a search bar contains "neuroglobin" and a "Display: Highlights" option. The main content area displays the entry for *605304, titled "NEUROGLOBIN; NGB". The entry includes a table of contents on the left with links to Title, Text, Cloning and Expression, Gene Structure, Mapping, Gene Function, Animal Model, References, Contributors, Creation Date, and Edit History. The main text area shows the HGNC Approved Gene Symbol: NGB, the Cytogenetic location: 14q24.3, and the Genomic coordinates (GRCh38): 14:77,265,490-77,271,311 (from NCBI). The entry also includes a section titled "TEXT" with a sub-section "Cloning and Expression" containing a paragraph about Hemoglobins and myoglobin.

*605304
Table of Contents

Title
Text
Cloning and Expression
Gene Structure
Mapping
Gene Function
Animal Model
References
Contributors
Creation Date
Edit History

* 605304
NEUROGLOBIN; NGB

HGNC Approved Gene Symbol: NGB

Cytogenetic location: 14q24.3 Genomic coordinates (GRCh38): 14:77,265,490-77,271,311 (from NCBI)

TEXT

▼ **Cloning and Expression**

Hemoglobins (e.g., HBA, 141800, and HBB, 141900) and myoglobin (MB; 160000) are members of the globin family, porphyrin-containing proteins that function in oxygen transport and storage in humans, other vertebrates, plants, fungi, and bacteria. Hemoglobins are tetrameric oxygen transporters in blood, whereas MB

...und wie komme ich zu meiner Sequenz?

Ich kenne eine Accession Number **NM_000518**

Ich kenne ein Gensymbol **HBB**
(Hämoglobin Beta)

Ich kenne einen passenden „Sequenz-Schnipsel“

mvhltpee**ks** **av**tal**wgkvn** **vdev**g**gealg** **rllv**vyp**wtg** **rf**fe**sfgdls** **tp**dav**mgnpk**
ag**tc**cttt**gg** **gg**atct**gtcc** **act**ctgat**g** **ctg**ttat**ggg** **ca**acc**ctaag** **gtg**aagg**ctc**

Sequenz-Datenbanken

NCBI	> GenBank (1979)
EBI	> EMBL database (1980) ENA European Nucleotide Archive
Genome-Net	> DDBJ = DNA database of Japan (1984)

- täglicher Abgleich erfolgt zwischen allen drei Datenbanken
- dennoch Unterschiede in der Redundanz und Annotations-Präzision

Probleme großer primärer Datenbanken

- REDUNDANZ: Archiv- alles bleibt drin
- alte u. z. T. falsche Einträge
- Vektorkontaminationen
- gleiche Seq - anderer Name
- inkonsistente Annotation

Lösung: „curated“ DBs



Probleme großer primärer Datenbanken

The screenshot shows the NCBI Entrez Gene search interface. The search bar contains 'ngb' and the results are displayed in a list format. The first result is for the gene 'Ngb' in *Mus musculus*, which has been replaced by GeneID: 69237. The second result is for 'NGB' in *Homo sapiens*, and the third is for 'Ngb' in *Mus musculus*. A black arrow points to the first result.

NCBI Entrez Gene

All Databases PubMed Nucleotide Protein Genome Struct

Search Gene for ngb Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 5 Send to

All: 50 Current Only: 41 Genes Genomes: 34 SNP GeneView: 14

Items 1 - 5 of 50

- ☐ 1: [Ngb](#)
GTP-binding protein NGB [*Mus musculus*]
Other Aliases: Crfg
Other Designations: chronic renal failure protein
GeneID: 85330
This record was replaced with [GeneID: 69237](#)
- ☐ 2: [NGB](#)
Official Symbol: NGB and **Name:** neuroglobin [*Homo sapiens*]
Chromosome: 14; **Location:** 14q24
MIM: 605304
GeneID: 58157
- ☐ 3: [Ngb](#)
Official Symbol: Ngb and **Name:** neuroglobin [*Mus musculus*]
Chromosome: 12; **Location:** 12 D3
GeneID: 64242

Protein-Sequenzdatenbanken

PIR-PSD („Protein Information ressource“ - „Protein Sequence Database“)

- größte öffentl. Proteindatenbank (250 000 Einträge)
- annotiert, nicht-redundant (?)
- > 2/3 der Sequenzen klassifiziert in 33 000 Super-Familien

Swiss-Prot (Amos Bairoch, Genf; jetzt vom EBI unterhalten)

- nicht-redundant, sehr informativ, äußerst exakt annotiert!
- Link zur PROSITE-Motivdatenbank (www.expasy.org/prosite)

PDB („Protein Data Bank“)

- bekannte 3D-Strukturen

NR • nicht-redundante Zusammenfassung von PIR, PDB, Swiss-Prot und allen aus GenBank-Nukleotid-Sequenzen übersetzten Proteinen !!

Protein-Sequenzdatenbanken

...die Konkurrenz zur NCBI-nr



UniProt
the universal protein resource

Home About UniProt Getting Started Searches/Tools Databases

Text Search BLAST Useful Tools/Links

Welcome to UniProt

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt has three components, each optimized for different uses. The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Reference Clusters (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via [text search](#), [BLAST similarity search](#), and [FTP](#).



[European Bioinformatics Institute](#)



[Swiss Institute of Bioinformatics](#)



[Georgetown University](#)

- www.uniprot.org
- EBI + PIR + Swissprot
- curation & annotation
- cross-referencing
- Such-Tools (Blast)

Protein-Sequenzdatenbanken

„abgeleitet“: Proteinfamilien, Domänen, Funktionelle Motive

The screenshot shows the EMBL-EBI InterPro website. At the top, there's a navigation bar with 'EMBL-EBI' and 'EB-eye Search' logos, followed by a search bar with 'All Databases' and 'Enter Text Here' fields, and 'Go', 'Reset', and 'Advanced' buttons. Below this is a menu with 'Databases', 'Tools', 'Groups', 'Training', 'Industry', 'About Us', and 'Help'. The main content area is titled 'Home' and contains a description of InterPro as a database of protein families, domains, and functional sites. It includes links to 'documentation', 'EBI Support', and a 'Search' button. A search bar with 'Search Entries' and 'Search InterPro' buttons is also present. On the left side, there's a sidebar with 'InterPro' and a list of links: 'InterPro home', 'InterProScan', 'Databases', 'Documentation', 'Tutorial', 'Project Outlines', 'Collaborators', 'Example Entry', 'Dataflow Scheme', 'Release Notes', 'User Manual', 'Publications', 'Browser FAQ', 'FTP site', 'Protein of the month', and 'Imports'. At the bottom, there's a section titled 'Announcement' with a bullet point about 'InterPro 14.1 is released'. Below that is an 'Information' section with several bullet points about E-values, match files, and UniParc matches. At the very bottom, there's a section titled 'InterPro Funding' with text about the grant number QLRI-CT-2000-00517 and the MRC-funded eFamily project. On the far left, there's a vertical stack of logos for UniProt, ProSite, Pfam, PRINTS, ProDom, and SMART.

- Integriert aus:
PROSITE, PRINTS, ProDOM,
PFAM, SMART, TIGRfam,
PIRSF, SUPERFAMILY,
Gene3D, Panther etc.

- >80 % aller Proteine in
UNIProt erfasst

- Text- und sequenzbasierte
Suche (<http://www.ebi.ac.uk/InterProScan/>)

DNA-Sequenzdatenbanken

(via NCBI)

GenBank

- 106 Milliarden Nukleotide (**Stand 2011! Bitte recherchieren..**)
- Größe verdoppelt sich alle 35 Monate
- ca. 1000 komplett sequenzierte Bakterien-Genome
- 380 Eukaryotengenome (in Arbeit)
- 67% der eingetragenen Sequenzen sind ESTs (1200 Spezies)
- > 300 000 Spezies repräsentiert (2200 neue/Monat)
- 12 % aller Sequenzen aus Mensch (13 Milliarden Bp)
- > 30 000 Zugriffe pro Tag
- GenBank ist in 17 Abteilungen unterteilt !

GenBank-Unterteilungen

Als *default* bei DB-Suchen meist eingestellt ist die

NR-nt

- nicht-redundante Zusammenfassung aus GenBank + EMBL + DDBJ + PDB
- Achtung: die Hochdurchsatz-Abteilungen EST/GSS/STS/HTG etc sind NICHT dabei!!

GenBank-Unterteilungen

Organismal Divisions:

Database	Division	BLAST	Example
BCT	Bacterial sequences	nr, month	Human Phase 3
PRI	Primate sequences	nr, month	
ROD	Rodent sequences	nr, month	
MAM	Other mammalian sequences	nr, month	
VRT	Other vertebrate sequences	nr, month	
INV	Invertebrate sequences	nr, month	Drosophila, C. elegans Phase 3
PLN	Plant and Fungal sequences	nr, month	Arabidopsis Phase 3
VRL	Viral sequences	nr, month	
PHG	Phage sequences	nr, month	
RNA	Structural RNA sequences	nr, month	
SYN	Synthetic and chimeric sequences	nr, month	
UNA	Unannotated sequences	nr, month	

Taxonomie

Functional Divisions:

Database	Division	BLAST	Example
EST	Expressed Sequence Tags	dbest, month	All Organisms: Phase 0, 1, and 2
STS	Sequence Tagged Sites	dbsts, month	
GSS	Genome Survey Sequences	dbgss, month	
HTG	High Throughput Genomic sequences	htgs, month	

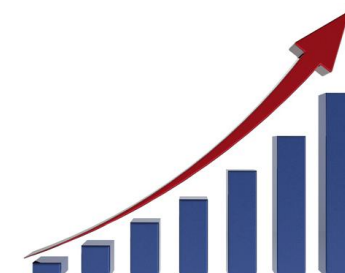
Hochdurchsatz-Daten

MONTH • neue Einträge der letzten 30 Tage aus GenBank/EMBL/DDBJ

Table 1.

Growth of GenBank divisions (nucleotide base-pairs)

Division	Description	Release 215 (August 2016)	Annual Increase (%)*
TSA	Transcriptome shotgun data	103 399 724 586	49.1%
WGS	Whole genome shotgun data	1 637 224 970 324	40.7%
BCT	Bacteria	26 474 028 571	36.9%
PHG	Phages	270 541 687	28.7%
PLN	Plants	14 705 679 094	22.9%
VRL	Viruses	2 973 938 989	19.2%
PRI	Primates	7 802 428 126	14.6%
PAT	Patent sequences	17 128 458 325	10.2%
UNA	Unannotated	204 984	9.3%
ENV	Environmental samples	5 218 628 157	7.7%
INV	Invertebrates	16 241 123 317	5.4%



RefSeq- „die Referenz unseres Wissensstandes“

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- ongoing curation by NCBI staff

RefSeq-

„die Referenz unseres Wissensstandes“

Table 1. RefSeq accessions, sequence type, processing method and categories

Accession format	Type	Method	Category
NC_123456	Genomic	Curated	Genomic molecules, available in Entrez Genomes (mitochondrion, viral and bacterial genomes, chromosomes)
NT_123456	Genomic	Assembled contigs	Genome annotation
NM_123456	mRNA	Computed	Predicted
		Curated	Provisional
		Curated	Reviewed
NG_123456	Genomic	Curated	Gene region
NP_123456	Protein	Computed; curated	Full-length proteins associated with curated nucleotide sequences
XM_123456	mRNA	Gene prediction	Genome annotation
XP_123456	Protein	Gene prediction	Genome annotation

- eine Sammlung **verifizierter** mRNAs, Gene und Proteine
- > 10 000 Organismen, > 11 Mio Gene/Proteine

GeneCards = Alternative zu NCBI

Location: <http://bioinfo.weizmann.ac.il/cards/index.html>

Google Elekt Zeitschr

GeneCards™ an academic web site of the **WEIZMANN INSTITUTE OF SCIENCE**

[Terms of Use](#) | [GeneCards Homepage](#) | [Search Examples](#) | [Comment Form](#)

Notice - Please read carefully prior to linking to any third-party site.

GeneCard for gene **CYGB**
GC17M074357

Approved [UCL/HGNC/HUGO Human Gene Nomenclature database](#) symbol
CYGB (cytoglobin)

Aliases and Additional Descriptions
(According to [GDB](#), [HUGO](#), and/or [SWISS-PROT](#))

- HGB
- STAP
- cytoglobin
- Cytoglobin (Histoglobin) (HGb) (Stellate cell activation-associated protein).

Chromosome: 17 [UDB/GeneLoc gene densities](#)

LocusLink cytogenetic band: 17q25.3 Ensembl cytogenetic band: 17q25.1

Gene in genomic location: bands according to Ensembl, locations according to [UDB/GeneLoc](#) (and/or [LocusLink](#) and/or [Ensembl](#) if different)

Chr 17

Chromosomal Location
(According to [UDB/GeneLoc](#) and/or [HUGO](#), and/or [LocusLink](#),
Genomic Views According to [UCSC](#) and [Ensembl](#))

Unified DataBase (GenBank)
Start: 74,357
End: 74,367
Size: 10,229
Orientation: minus

Unified DataBase (GenBank)
Genomic View:
[UCSC Genome Browser](#)

CYGB expression in normal human tissues based on quantifying ESTs from various tissues in Unigene clusters (Build 155 Homo sapiens).

Tissue	Clones per gene	Total clones
BMR Bone marrow	0	26,809
SPL Spleen	0	13,489
TMS Thymus	0	3,451
BRN Brain	20	274,393
SPC Spinal cord	0	506
HRT Heart	3	35,078
MSL Skeletal muscle	0	23,264
LVR Liver	1	55,430
PNC Pancreas	0	58,927
PST Prostate	1	81,135
KDN Kidney	1	121,315
LNG Lung	2	167,397



Für Entdecker!

Die Sequenzen hier sind meist unpubliziert...

- **Trace Archive:** unannotierte Sanger-Reads aus EST-und Gesamtgenom-Projekten (>500 Spezies, 2.1 Milliarden Reads)
- **Short Read Archive:** NGS reads diverser Technologien (TBp!!!)

Last week Top 10 Arrivals (10/25/2009 - 10/31/2009)	
Organism	Count
HOMO SAPIENS	1,819,859
HUMAN GUT METAGENOME	595,201
SUS SCROFA	113,473
HUMAN METAGENOME	55,242
CHRYSEMYS PICTA	43,916
BODO SALTANS	16,060
CAVIA PORCELLUS	13,048
OCHOTONA PRINCEPS	10,486
MOUSE GUT METAGENOME	8,739
CALLITHRIX JACCHUS	8,453

Genom-Browser 1

www.ensembl.org

[illegible]

- > 100 Genome, manche nur hier zu finden...

Genom-Browser 2

<http://genome.ucsc.edu/>

The screenshot shows the UCSC Genome Browser Gateway. At the top is a navigation bar with links: Home, Genomes, Blat, Tables, Gene Sorter, PCR, Session, FAQ, Help. Below this is the title "Human (*Homo sapiens*) Genome Browser Gateway". A paragraph states: "The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#). Software Copyright (c) The Regents of the University of California. All rights reserved." Below this is a search form with five fields: "clade" (dropdown menu showing "Vertebrate"), "genome" (dropdown menu showing "Human"), "assembly" (dropdown menu showing "Mar. 2006"), "position or search term" (text input field containing "chr17:44,064,851-44,064,920"), and "image width" (text input field containing "620"). There is a "submit" button to the right of the "image width" field. Below the search form is a link: "Click [here to reset](#) the browser user interface settings to their defaults." Below this link are three buttons: "add custom tracks", "configure tracks and display", and "clear position". Below the search form is a section titled "About the Human Mar. 2006 (hg18) assembly ([sequences](#))". The text below this title says: "The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium". Below this is a section titled "Sample position queries". The text below this title says: "A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid positions for the human genome. See the [User's Guide](#) for more information." Below this is a table with two columns: "Request:" and "Genome Browser Response:". The first row shows "chr7" in the "Request:" column and "Displays all of chromosome 7" in the "Genome Browser Response:" column.

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade genome assembly position or search term image width

Vertebrate Human Mar. 2006 chr17:44,064,851-44,064,920 620 submit

[Click here to reset](#) the browser user interface settings to their defaults.

[add custom tracks](#) [configure tracks and display](#) [clear position](#)

About the Human Mar. 2006 (hg18) assembly ([sequences](#))

The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid positions for the human genome. See the [User's Guide](#) for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7

- derzeit:
46 Säuger
+ Invertebraten
usw.
- schnell, inter-
aktiv, flexibel
- LinkOuts zu
div. sekund. DBs

Genom-Browser 3

<http://www.ncbi.nlm.nih.gov/Genomes/>

The screenshot displays the NCBI Map Viewer interface. At the top, the NCBI logo is on the left, and the 'NCBI Map Viewer' title with a compass icon is on the right. Below the title is a navigation bar with tabs: PubMed, Nucleotide, Protein, Genome, Gene, Structure, Pop Set, and Taxonomy. The 'Genome' tab is selected. A search bar contains the text 'Search for' followed by a text input field, 'on chromosome(s)' followed by another text input field, and a dropdown menu set to 'All'. A 'Find' button is to the right of the dropdown. On the left side, a blue sidebar contains a 'Map Viewer' section with links: Map Viewer Home, Map Viewer Help, Human Maps Help, and Release Notes. Below this is an 'NCBI Resources' section with links: Genome Project, TaxPlot, Consensus Coding Sequence (CCDS), Human Genome Resources, NCBI Handbook, RefSeq, and Whole Genome Association (WGA). The main content area is titled 'Homo sapiens (human) genome view' and includes links for 'Build 36.2 statistics' and 'Switch to previous build'. A link for 'BLAST search the human genome' is also present. The genome is visualized as two rows of chromosome bars. The top row shows chromosomes 1 through 13, and the bottom row shows chromosomes 14 through 22, X, Y, and MT. Each chromosome is represented by a black bar with a white band. Below each bar is a number or label: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y, MT. At the bottom, a blue box contains the 'Lineage' information: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo; Homo sapiens.

NCBI

NCBI Map Viewer

PubMed Nucleotide Protein Genome Gene Structure Pop Set Taxonomy

Search for on chromosome(s) assembly All

Map Viewer

- Map Viewer Home
- Map Viewer Help
- Human Maps Help
- Release Notes

NCBI Resources

- Genome Project
- TaxPlot
- Consensus Coding Sequence (CCDS)
- Human Genome Resources
- NCBI Handbook
- RefSeq
- Whole Genome Association (WGA)

Homo sapiens (human) genome view

Build 36.2 statistics [Switch to previous build](#)

[BLAST search the human genome](#)


1 2 3 4 5 6 7 8 9 10 11 12 13

14 15 16 17 18 19 20 21 22 X Y MT

Lineage: [Eukaryota](#); [Metazoa](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Euteleostomi](#); [Mammalia](#); [Eutheria](#); [Euarchontoglires](#); [Primates](#); [Haplorhini](#); [Catarrhini](#); [Hominidae](#); [Homo](#); [Homo sapiens](#)

Genom-Browser


- Humangenom: Assembly selbst nur noch am NCBI
- Humangenom: alle 3 Browser zeigen „NCBI build“, aber u. U. unterschiedliche Versionen
- nur UCSC zeigt alte builds
- gezeigte andere Spezies differieren !
- unterschiedliche Datenquellen und Methoden für Annotation!!




FlyBase

[Report A Bug](#)


Home Tools Files Species Documents Resources News Help Archives




BLAST




GBrowse



QueryBuilder



TermLink



ImageBrowse

News

[Profile Manager released](#) | Feb 07
[12 genomes - publication plans](#) | Dec 06
[Tucson - new prices](#) | Dec 06

Upcoming Meetings

[48th Drosophila Conference](#) | 7 Mar 07
[8th Int. Dros. Heterochromatin](#) | 3 Jun 07
[8th Japanese Dros. Res. Conf.](#) | 2 Jul 07
[3rd Int. Mosquito/Vector Meet.](#) | 12 Jul 07

Quick Search

Search:

Data Class:

Enter text:

☒ Dmel only
☐ All species

☒ ID/Symbol/Name
☐ All text

genes

[Find A Fly Person](#)

[QuickSearch help](#)

FASTA-Flatfile-Format

■ 1: NP_071859. neuroglobin [Mus ...[gi:11967939]

```
>gi|11967939|ref|NP_071859.1| neuroglobin [Mus musculus]
MERPESELIQSWRVVSRSPLEHGTVLFARLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLTSLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSRLYGAV
VQAMSRGWDGE
```

■ 1: AJ245945. Mus musculus mRNA...[gi:10639821]

```
>gi|10639821|emb|AJ245945.1|MMU245945 Mus musculus mRNA for neuroglobin (Ngb gene)
GCTGCATGTGCGTTGACTGCACCCACGCCTCGAGGGTCCCATCACTGCGTCCCGCGAGTCTCCTGGGAGA
GAGAGCATGGAGCGCCCGGAGTCAGAGCTGATCCGGCAGAGCTGGCGGGTAGTGAGCCGCAGCCCTCTGG
AACATGGCACTGTCCTGTTCCGCCAGGCTCTTCGCCCTGGAACCCAGCCTGCTGCCTCTCTTCCAGTACAA
TGGCCGCCAGTTCTCCAGCCCTGAGGACTGTCTCTCCTCTCCAGAATTCCTGGACCACATTAGGAAGGTG
ATGCTAGTGATTGATGCTGCAGTGACCAACGTGGAGGACCTGTCTTCATTGGAGGAGTACCTGACCAGCT
TGGGCAGGAAGCATCGGGCAGTGGGAGTGAGGCTCAGCTCCTTCTCGACAGTAGGCGAGTCCCTGCTCTA
CATGCTGGAGAAGTGCCTGGGTCCCGACTTTACACCAGCTACAAGGACCGCCTGGAGCCGACTCTACGGA
GCTGTGGTGCAAGCCATGAGCCGAGGCTGGGATGGGGAGTAAGAGACGAGCCAGTGCCCCCTATCTATGTG
TGTCTGTCTGTTGATCTGCCTGTTGTAGTCTTAGCCTCTCCCCAGGGTCTCTCTATACCTTGCTC
```

...das FASTA-Format kann von vielen Sequenzverarbeitungs-
Programmen problemlos gelesen werden

Datenbanken und Computer-Tools arbeiten mit unterschiedlichen Sequenzformaten

- | | |
|--------------------|-----------------------|
| 1. IG/Stanford | 10. Olsen (in-only) |
| 2. GenBank/GB | 11. Phylip3.2 |
| 3. NBRF | 12. Phylip |
| 4. EMBL | 13. Plain/Raw |
| 5. GCG | 14. PIR/CODATA |
| 6. DNASTrider | 15. MSF |
| 7. Fitch | 16. ASN.1 |
| 8. Pearson/Fasta | 17. PAUP/NEXUS |
| 9. Zuker (in-only) | 18. Pretty (out-only) |

Lösung: Programme wandeln Formate um!

READSEQ <https://www.ebi.ac.uk/Tools/sfc/readseq/>

Seqverter <http://www.genestudio.com/seqverter>