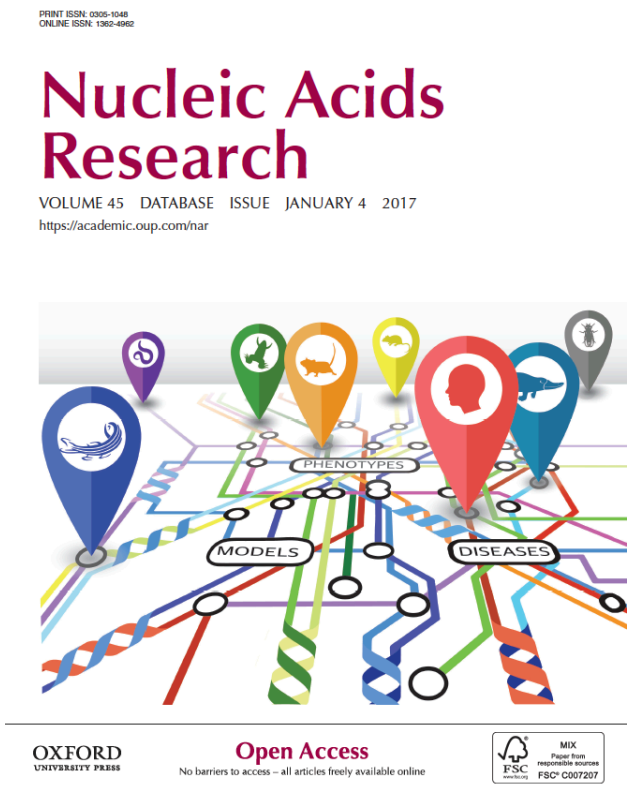
WS2017/2018

F1-Praktikum
„Genomforschung und Sequenzanalyse
- Einführung in Methoden der Bioinformatik- “



Datenbanken in der Molekularbiologie

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb,
ARR, AsDb, BBDB, BCGD, Beanref, BioImage,
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,
Genline, 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, 0-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… !!!!

Folie: Tal Dagan, D‘dorf

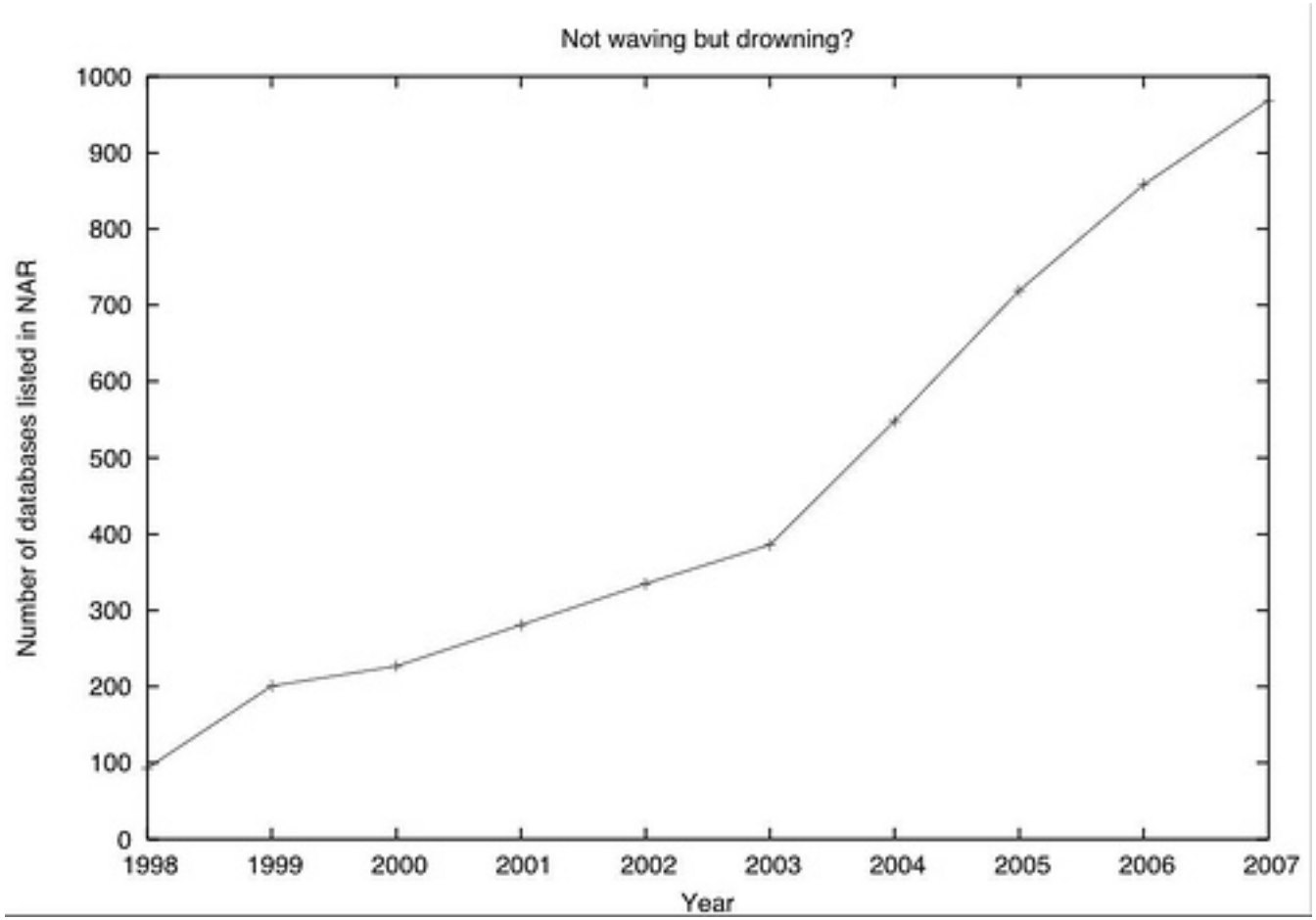
Datenbanken in der Molekularbiologie

- >> 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

z. B.

„Genbank“

„Flybase“

„Wanda“: a library of duplicated fish genes“

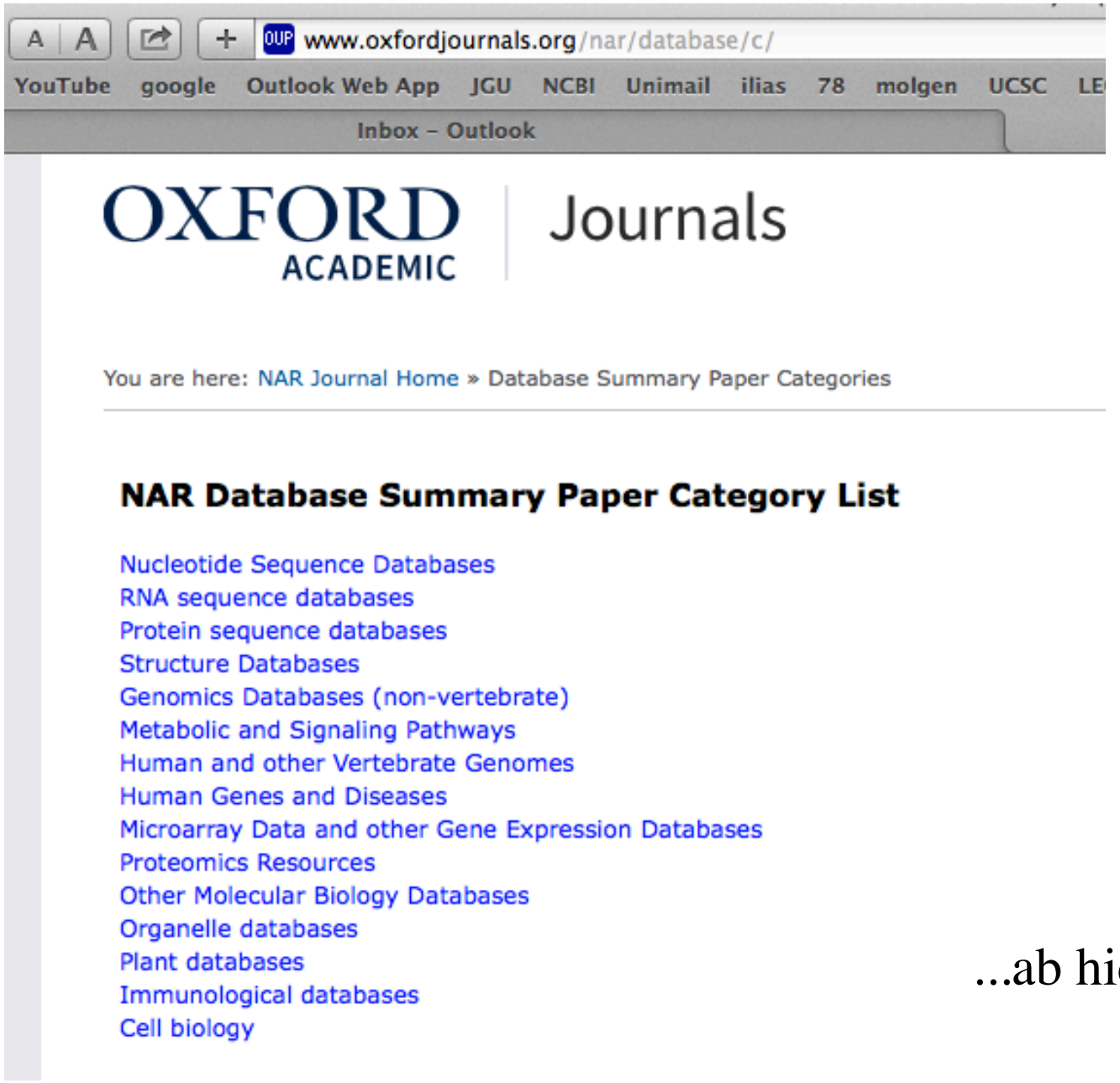
„ENZYME“

- 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

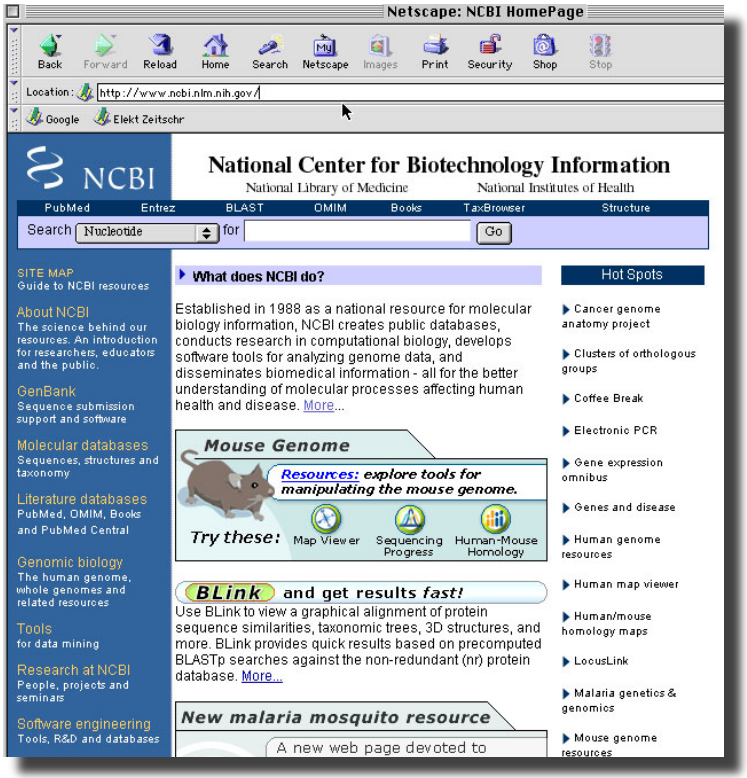


...ab hier weiter suchen

Datenbanken in der Molekularbiologie

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

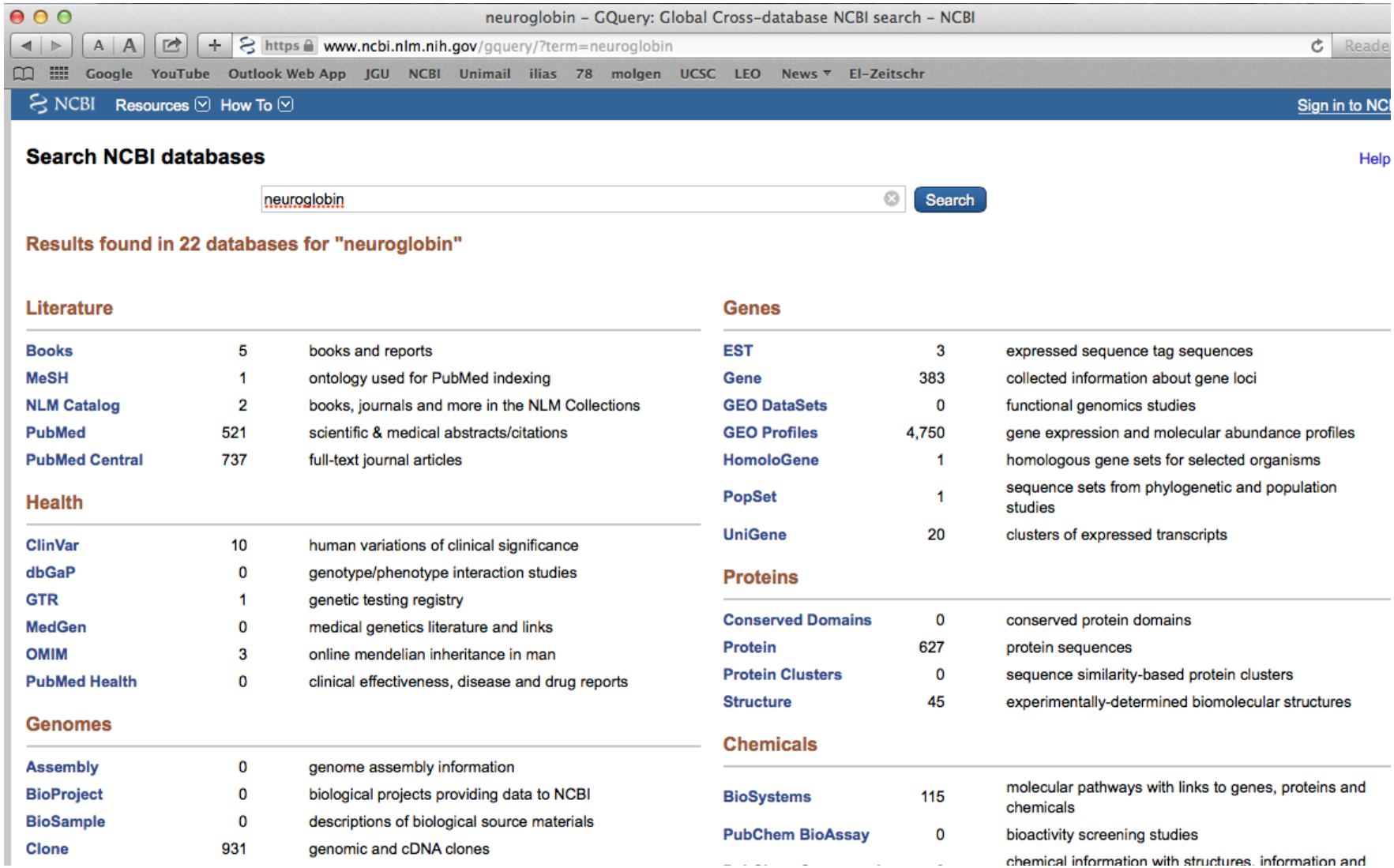
National Center for Biotechnology Information, Am NIH, Bethesda, Maryland, USA



http://www.ebi.ac.uk

European Bioinformatics Institute, Sanger Campus, Hinxton, GB

Integrierte Such-Werkzeuge!



www.ncbi.nlm.nih.gov/

Der Einstieg in die Suche...

watson jd

Myoglobin

j mol evol

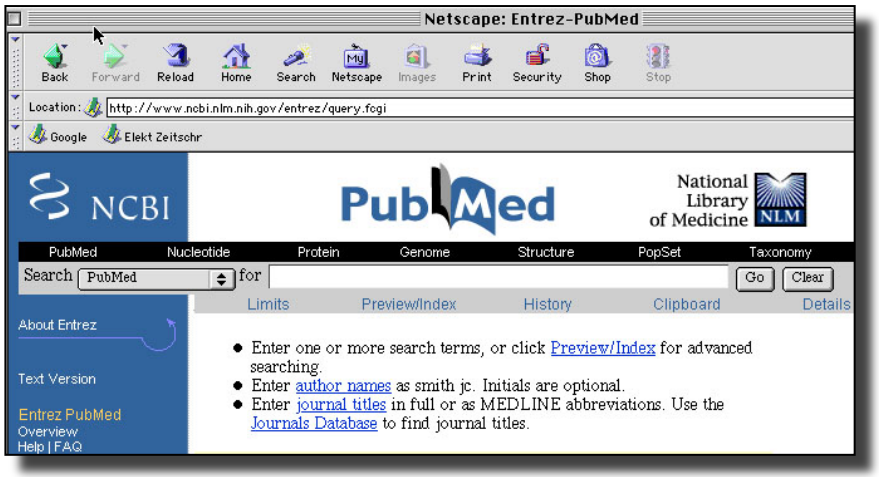
3M syndrome

P02998

mouse gtpase

Exkurs....

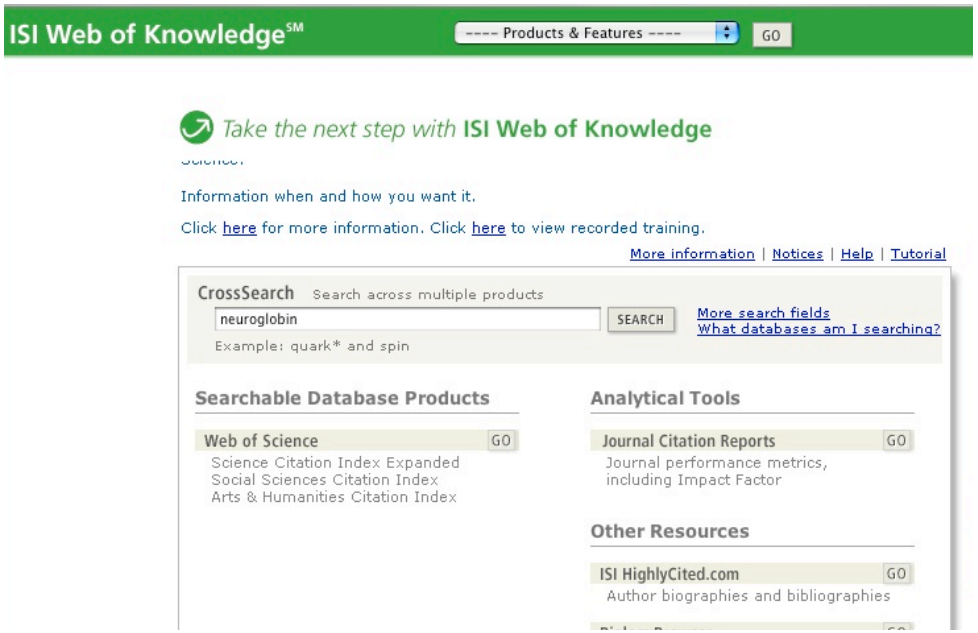
Literatur-Suche



- 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



- 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 citated 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: Globins (replace with **descriptive** term)

search term	search field	connector
neuroglobin		AND
cytoglobin		AND
globin AND developme		AND
development AND hypc		AND

Save! Reset!

Type: PubMed PubMed Neighbour Nucleotide Nucleotide Neighbour

Terms: 5 Change!

Finish!

- durchsucht PubMed und ENTREZ-DBs > Alarm!!

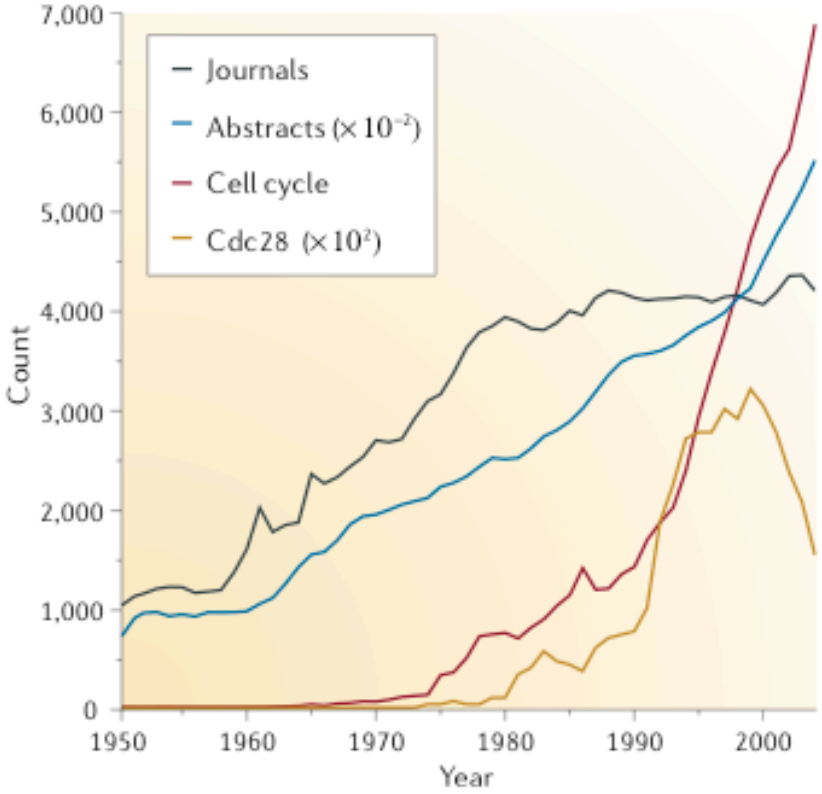


Figure 1 | **Growth of Medline.** The numbers of journals, papers (as represented by Medline abstracts), papers on the cell cycle and papers on Cdc28 that were published each year from 1950 to 2005 are shown. An average for 3 years was calculated for the Cdc28 curve because of much lower numbers. The 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.

Literatur-Suche

Zunehmend
ein Problem…

Lit-Suche
> Discovery

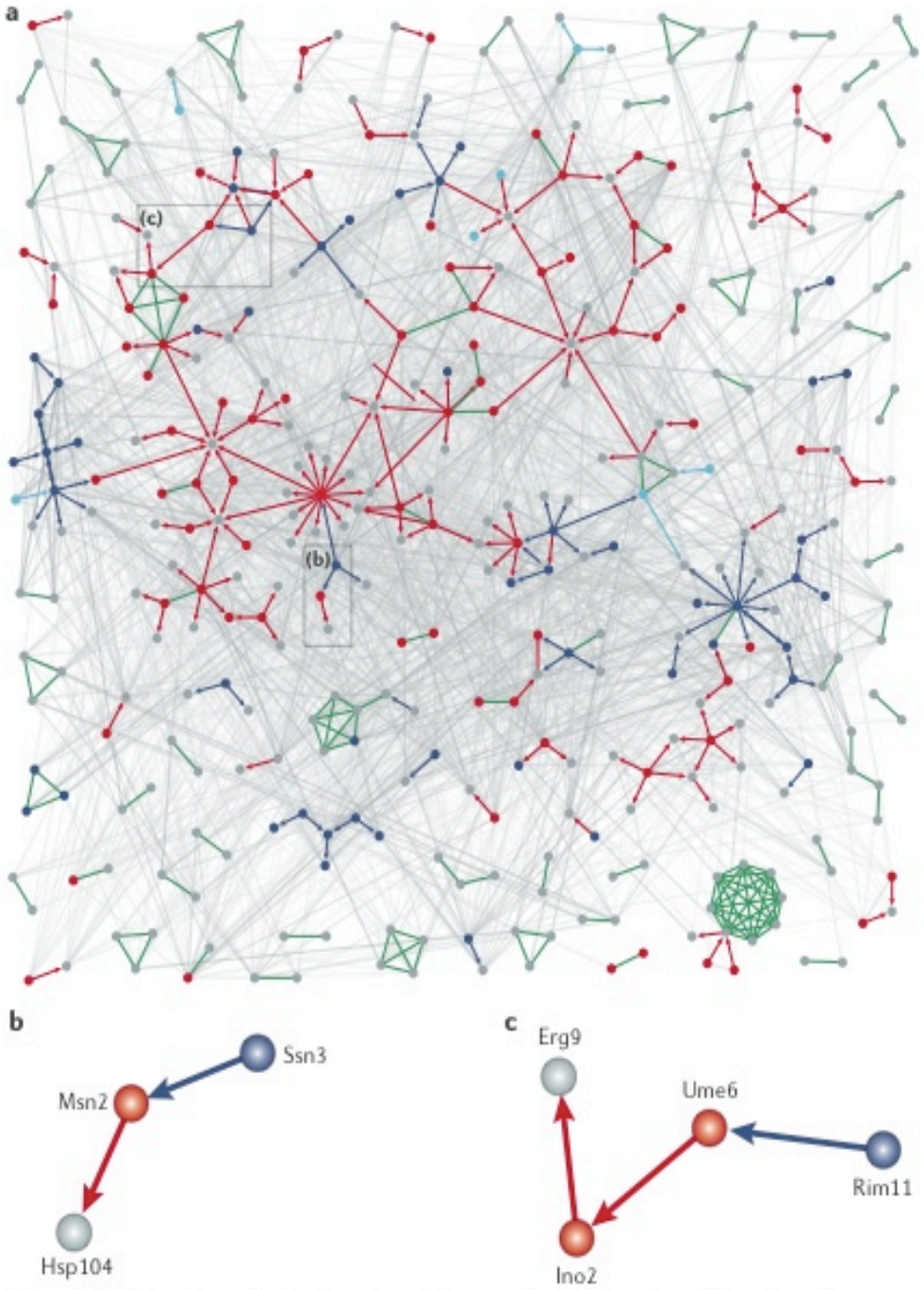
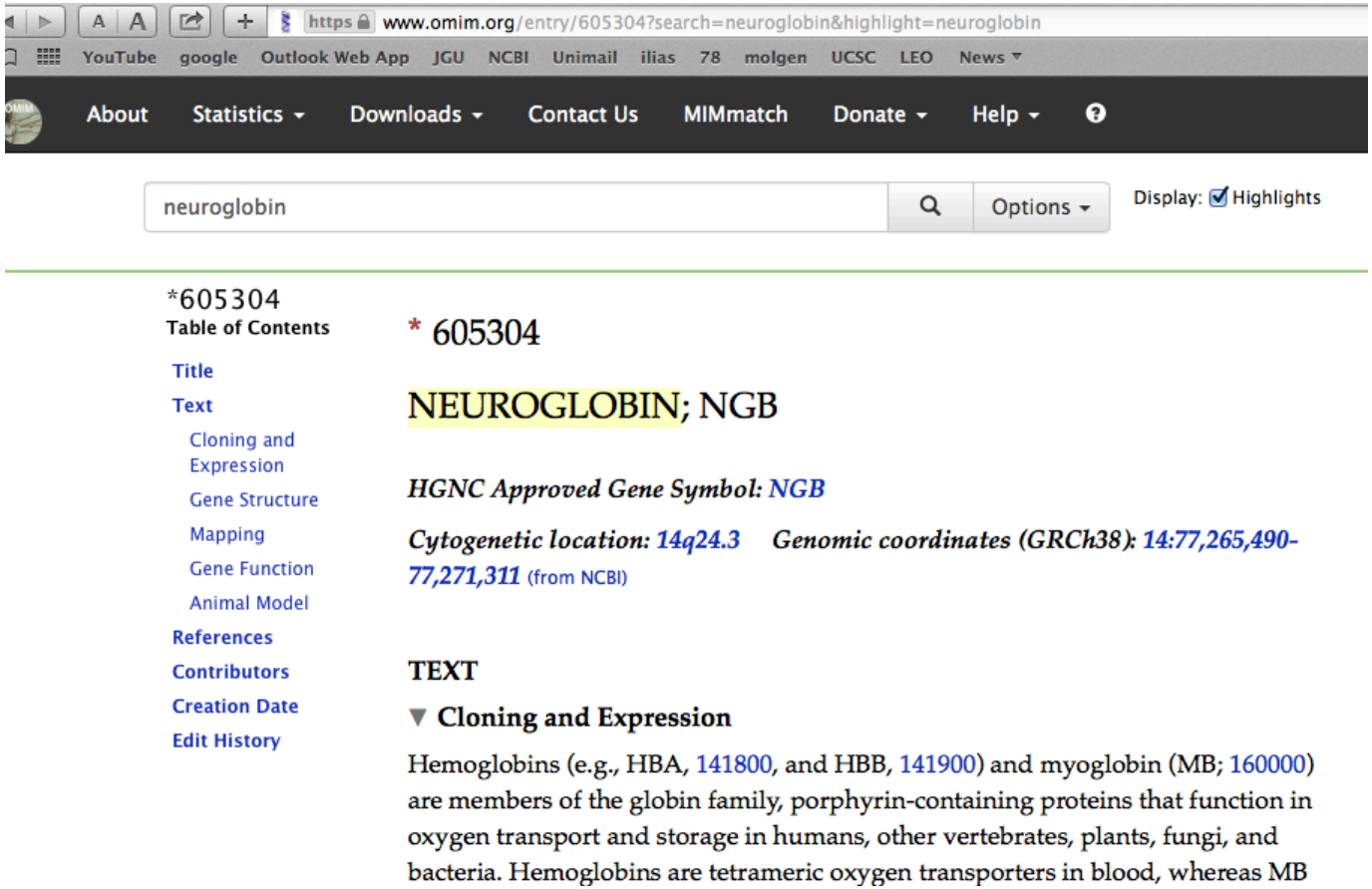


Figure 3 | **A literature-derived network for yeast. a** | A yeast protein network was derived that applied information-extraction approaches to all abstracts that are stored in Medline, using both a statistical co-occurence method[54] and a natural-language-processing (NLP)-based one[62]. Functional 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



…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“

```
mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
agtcctttgg ggatctgtcc actcctgatg ctgttatggg caaccctaag gtgaaggctc
```

Sequenz-Datenbanken

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

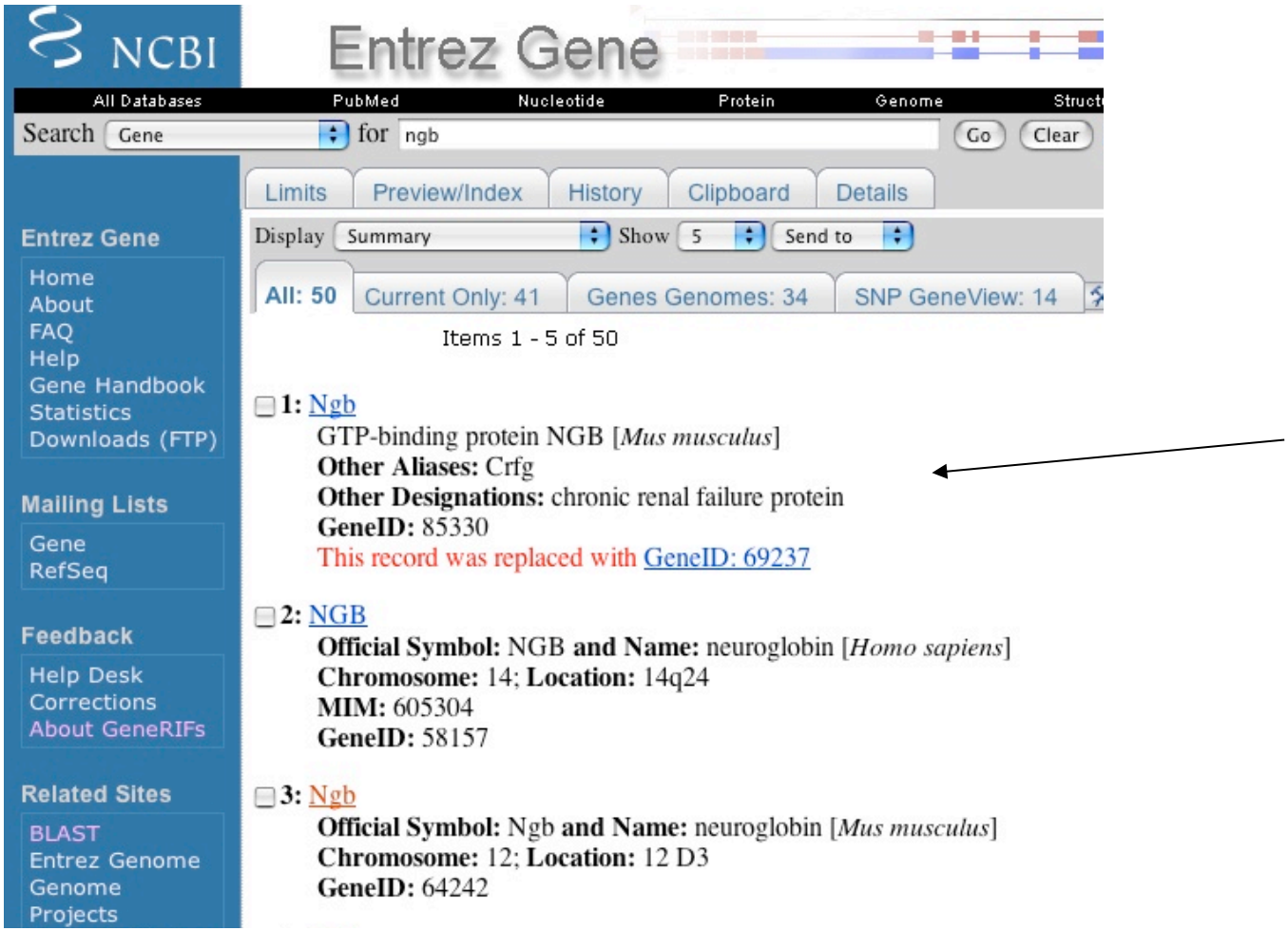
- täglicher Abgleich erfolgt zwischen allen drei Datenbanken
- dennoch Unterschiede in der Redundanz und Annotations-Präzision

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



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 !!

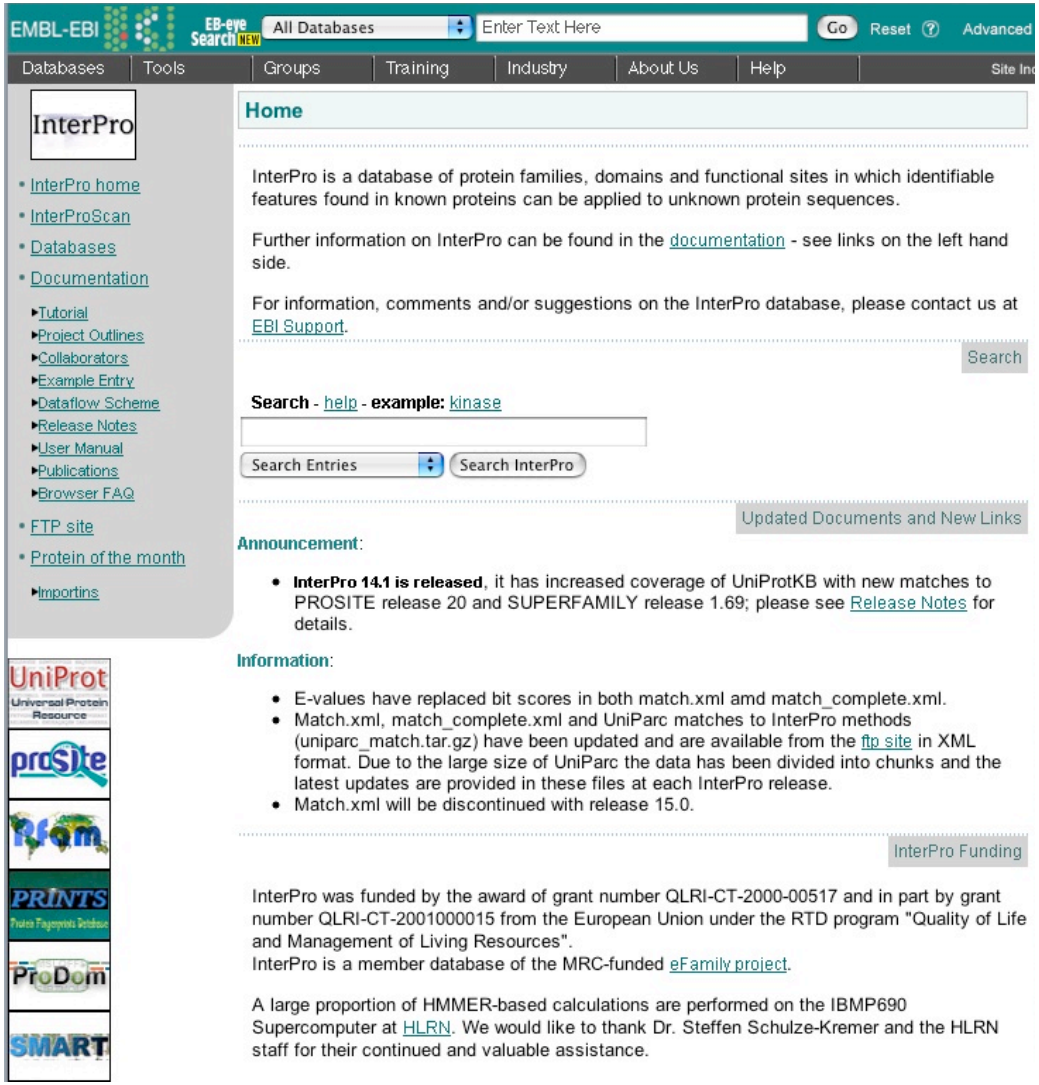
Protein-Sequenzdatenbanken

...die Konkurrenz zur NCBI-nr

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

Protein-Sequenzdatenbanken

„abgeleitet“: Proteinfamilien, Domänen, Funktionelle Motive



- 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	
PRI	Primate sequences	nr, month	Human Phase 3
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	
STS	Sequence Tagged Sites	dbsts, month	
GSS	Genome Survey Sequences	dbgss, month	
HTG	High Throughput Genomic sequences	htgs, month	All Organisms: Phase 0, 1, and 2

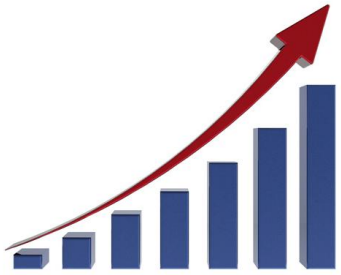
Hochdurch-satz-Daten

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%



www.ncbi.nlm.nih.gov/RefSeq/

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

www.ncbi.nlm.nih.gov/RefSeq/

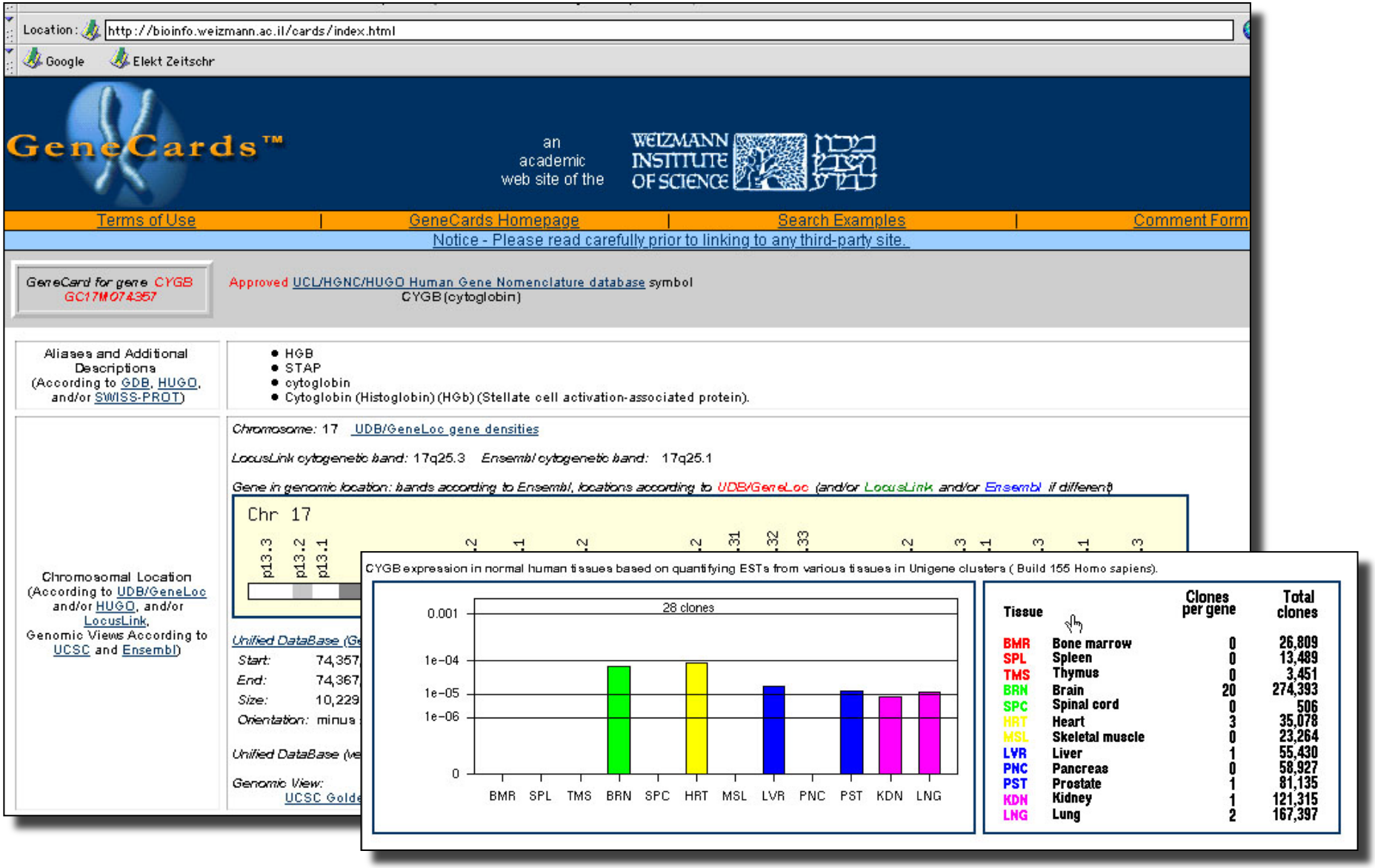
RefSeq-
„die Referenz unseres Wissensstandes“

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



Für Entdecker!

Die Sequenzen hier sind meist unpubliziert...

- Trace Archive: unannotierte Sanger-Reads aus EST-und Gesamtgenom-Projekten (>500 Spezies, 2.1 Milliarden Reads)

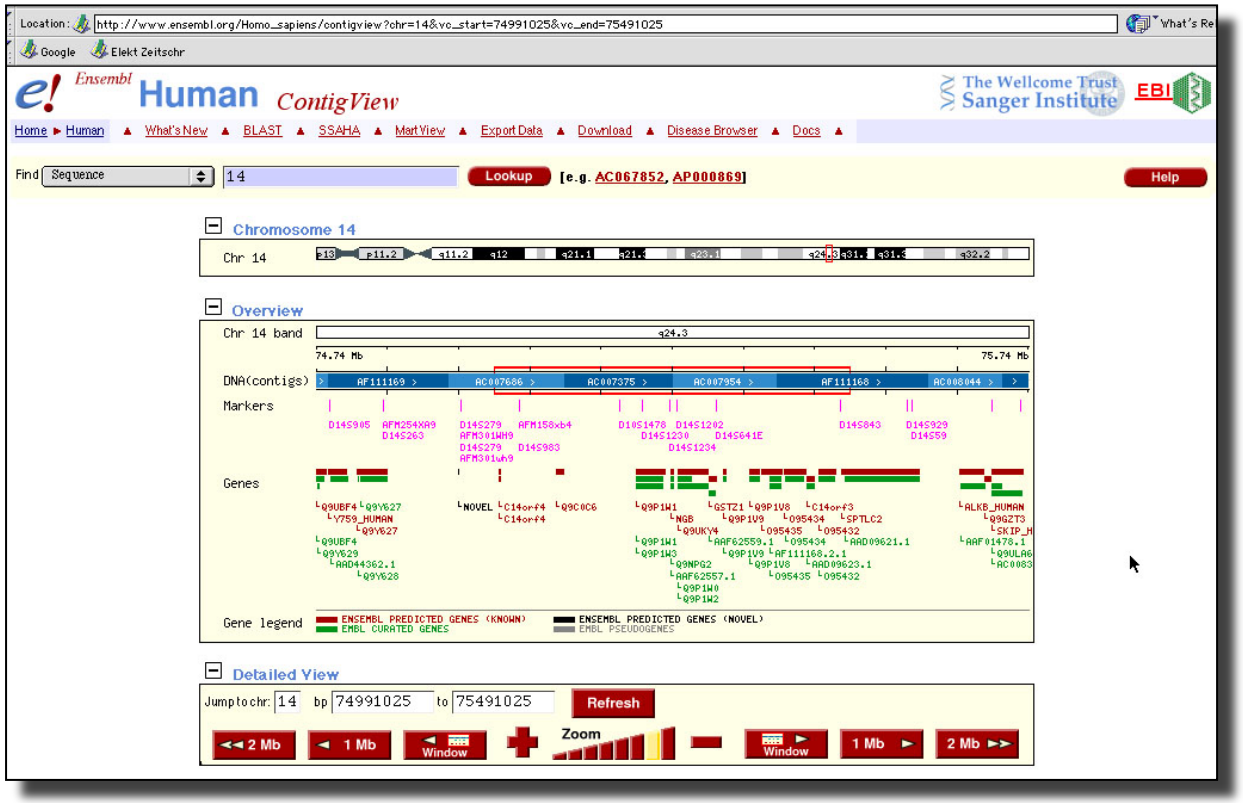
- Short Read Archive: NGS reads diverser Technologien (TBp!!!)

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

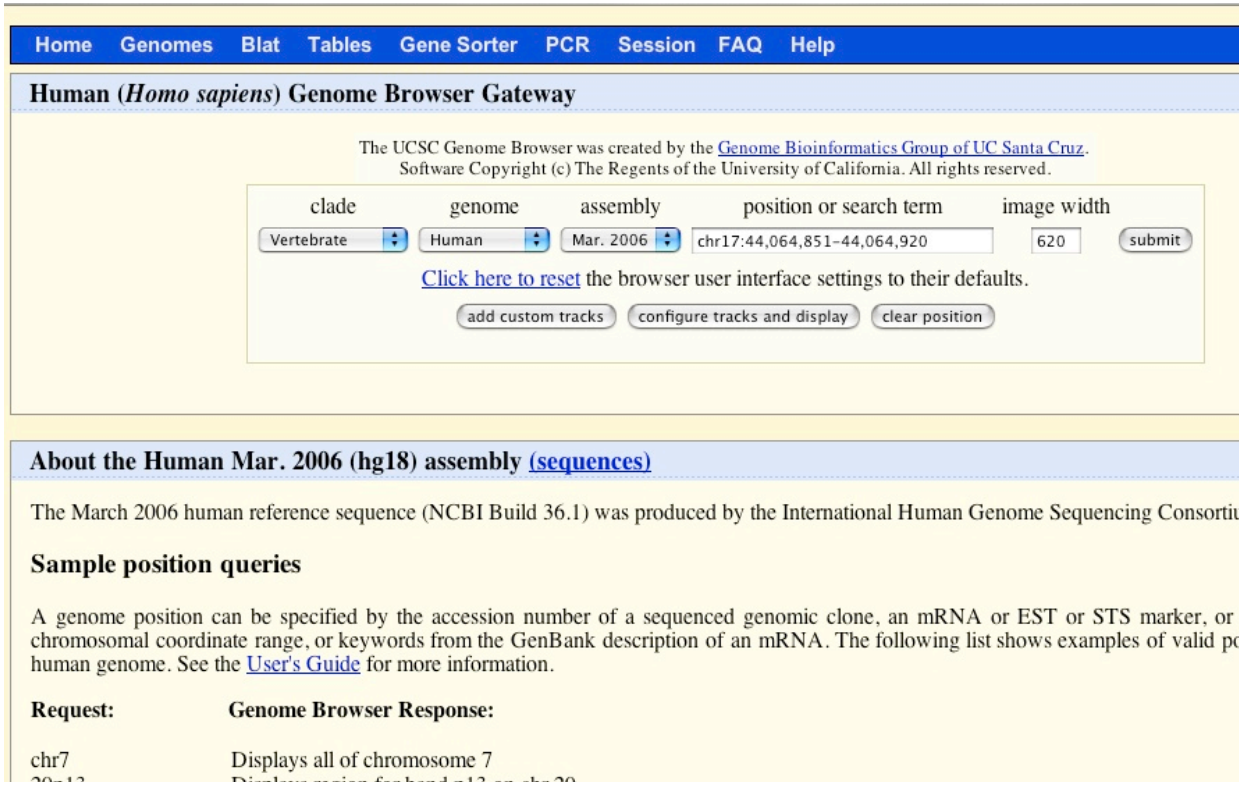
www.ensembl.org



- > 100 Genome, manche nur hier zu finden…

Genom-Browser 2

http://genome.ucsc.edu/



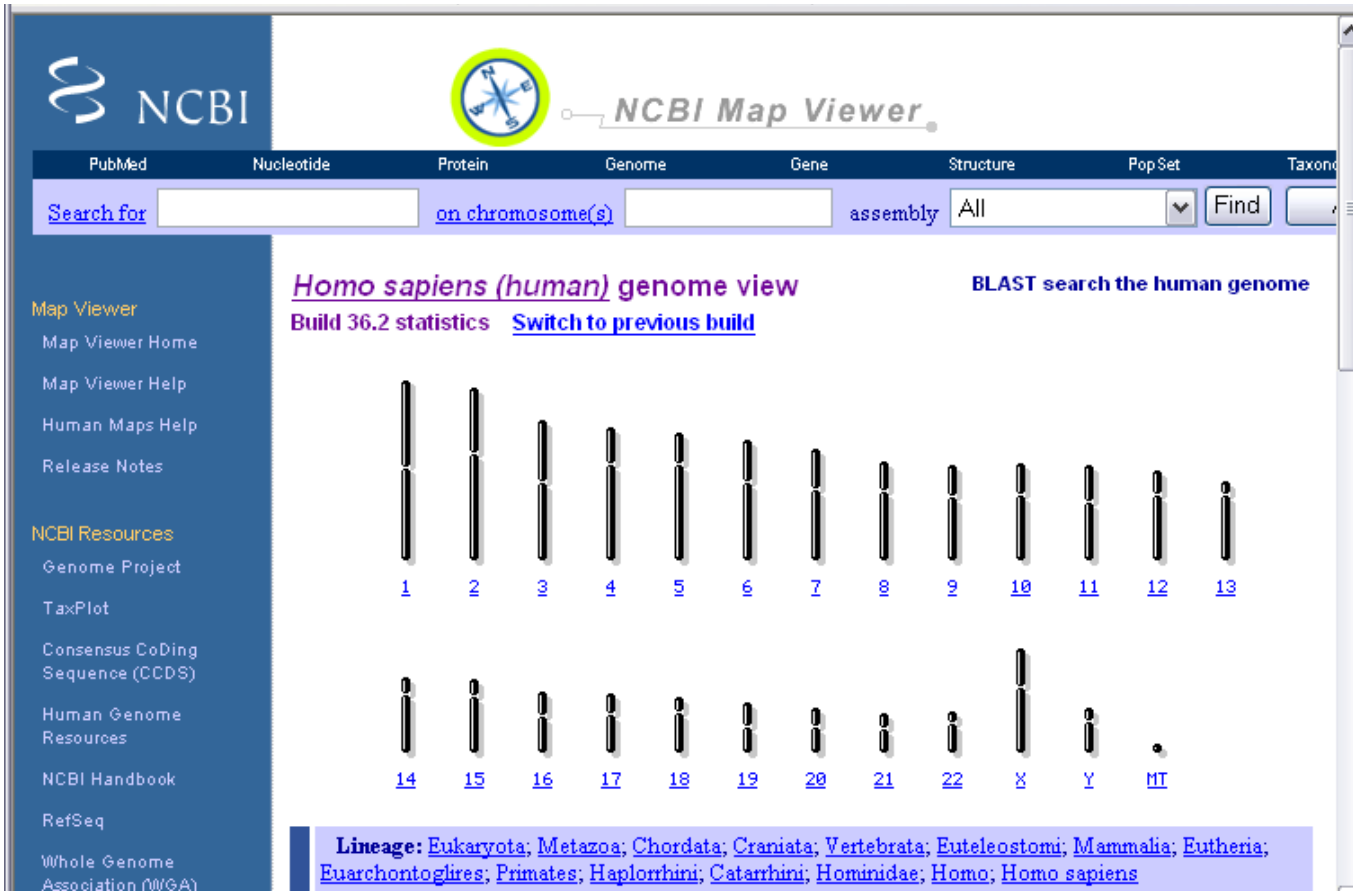
- derzeit:
  46 Säuger
  + Invertebraten
  usw.

- schnell, inter-
  aktiv, flexibel

- LinkOuts zu
  div. sekund. DBs

Genom-Browser 3

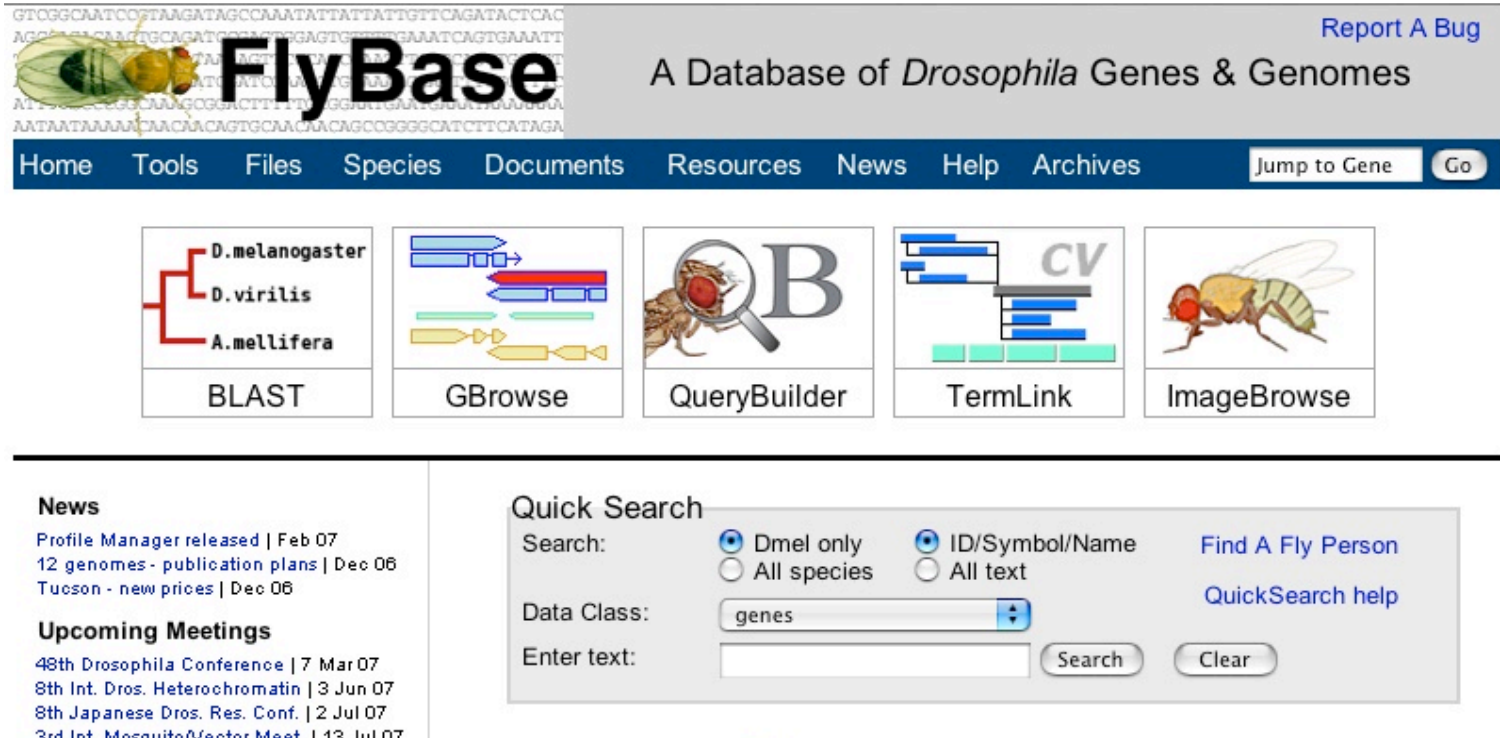
http://www.ncbi.nlm.nih.gov/Genomes/



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!!

Spezielle Genom-Browser



http://flybase.bio.indiana.edu/

FASTA-Flatfile-Format

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

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

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

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

...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