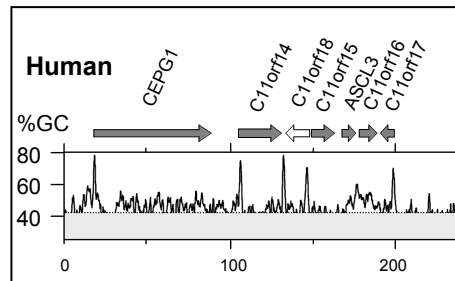


WS 2018/2019

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

Thomas Hankeln



Strategien der Gensuche

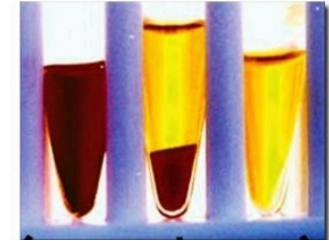


Datenbanken in der Molekularbiologie

Strategien der Gensuche

- der **biochemische** Weg („funktionales Klonieren“)
- der **genetische** Weg („positionelles Klonieren“)
- der „**Genomics**“-Weg

Der biochemische Weg



Voraussetzung: Das Genprodukt (Protein) muss bekannt sein!

Reinigung des Proteins



Antikörperherstellung

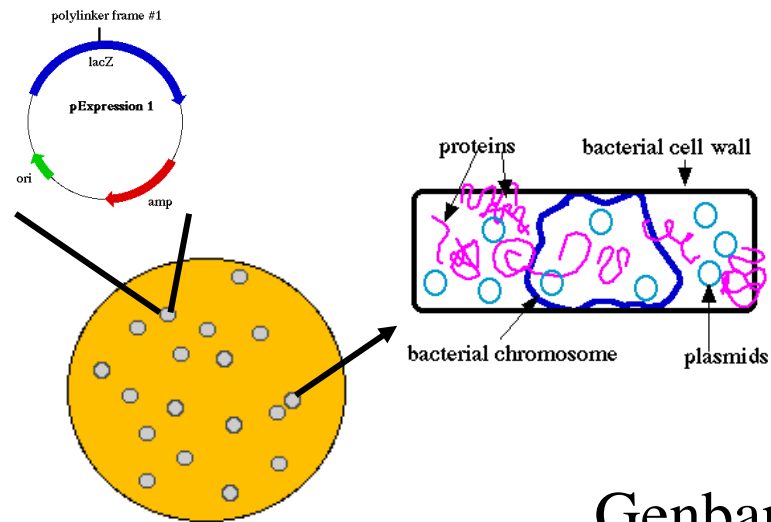


„Screenen“ einer cDNA-Expressions-
klonbank mit **AK als Sonde**

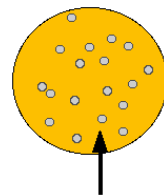


Identifizierung des Klons, der die cDNA
des gesuchten Gens enthält und daher das
gesuchte Protein bildet

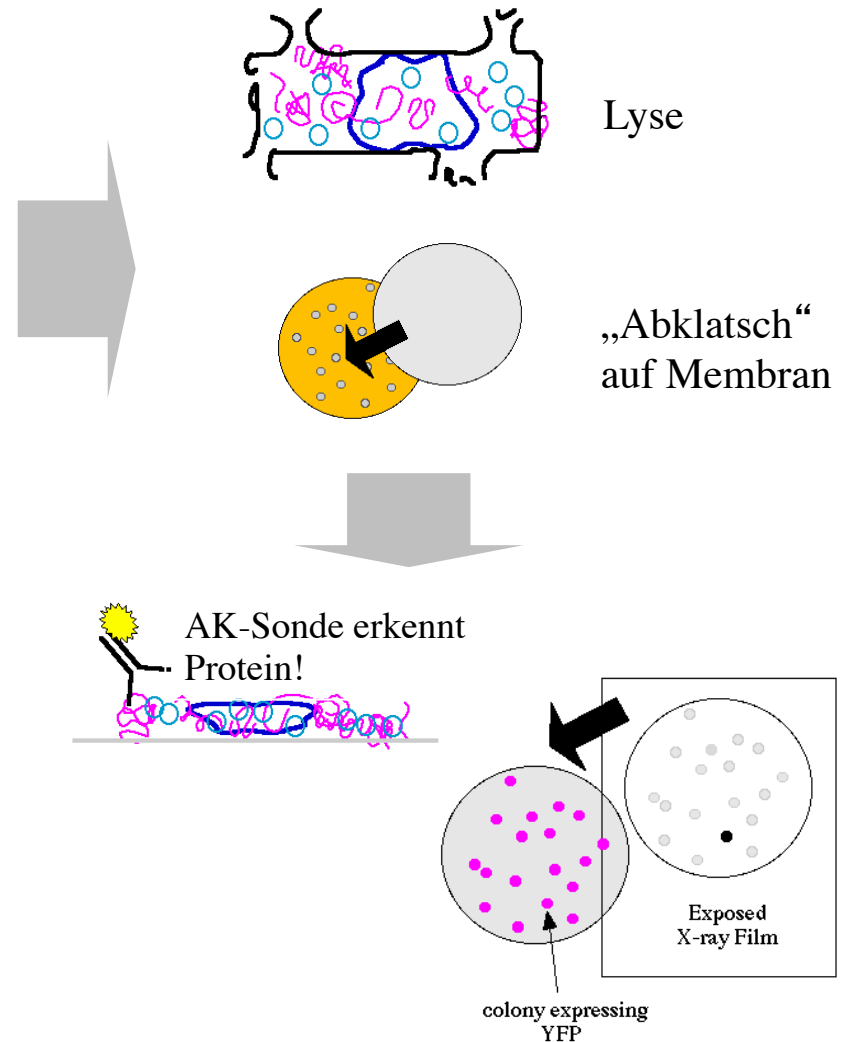
Der biochemische Weg



Genbank:
jeder Bakterienklon enthält und exprimiert ein
anderes Gen



**Der Klon enthält
das gesuchte Gen!!**



Der genetische Weg:

Positionelle Klonierung

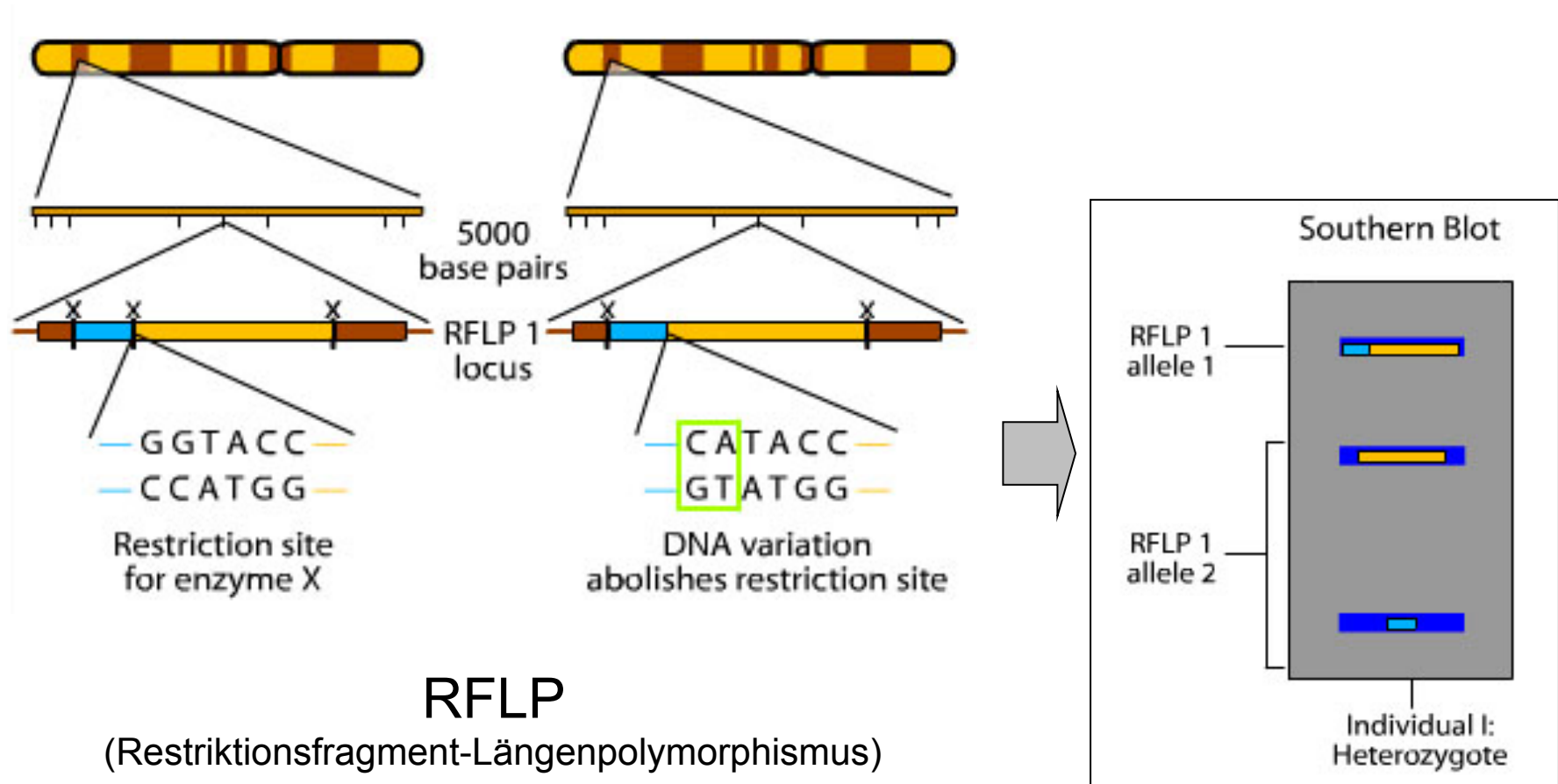
Voraussetzung:

Information über die Lage eines (Krankheits)gens auf den Chromosomen bzw. im Genom

- sichtbare Chromosomen-Anomalien?
- Verlust der Heterozygotie (LOH) in einer Genomregion?
(>Tumor-Suppressorgene)
- **Kopplungsanalyse:**
Wird meine Erkrankung (in großen Familien) zusammen mit leicht unterscheidbaren „genetischen Markern“ vererbt (deren Position ich bereits kenne) ?

Positionelle Klonierung braucht Marker

„Marker“ sind DNA-Loci, an denen es **Unterschiede** zwischen Genomen (Allele) gibt, deren Vererbung verfolgt werden kann



weitere gebräuchliche Markertypen

- **Mikrosatelliten** > Detektion per PCR & Gelelektrophorese

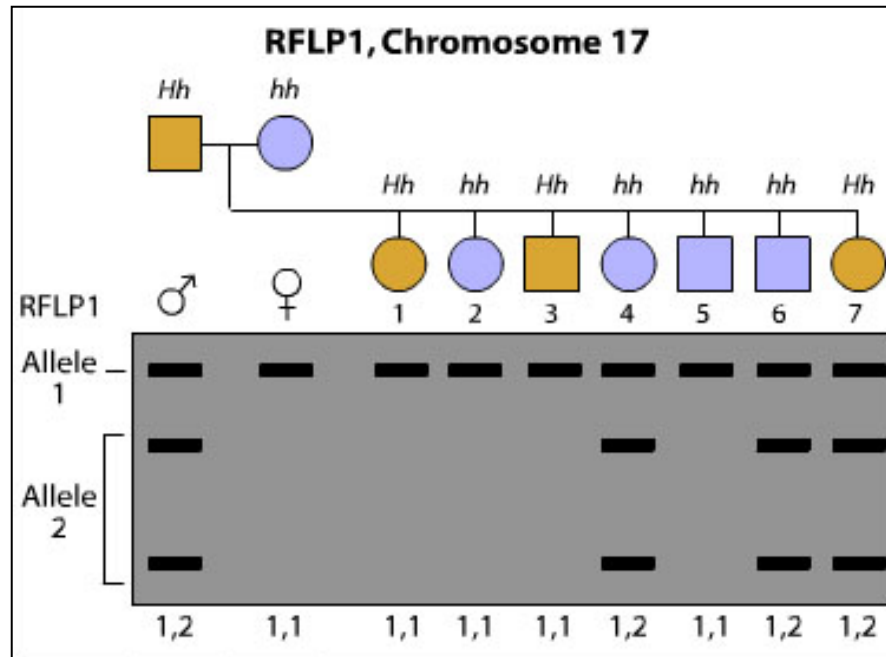


- **SNPs** single nucleotide polymorphisms >Detektion per PCR & Sequenzierung

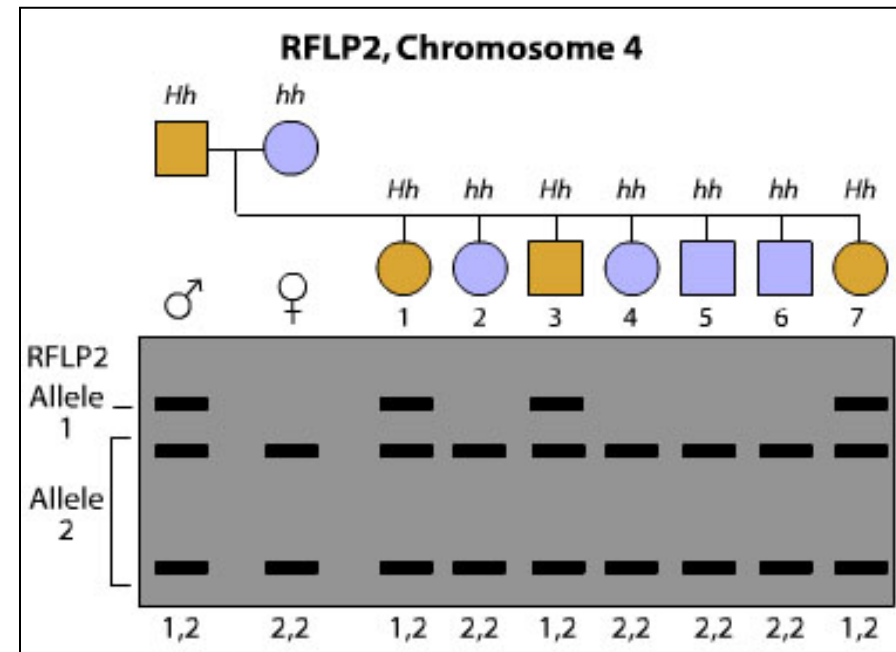


Positionelle Klonierung zeigt...

Wo ist das Huntington-Gen?



Krankes Allel wird nicht
zusammen mit Allel 2 vererbt
> Gen liegt nicht auf Chr. 17!



Huntington-Allel wird
zusammen mit Allel 1 vererbt
> Gen liegt auf Chr. 4!

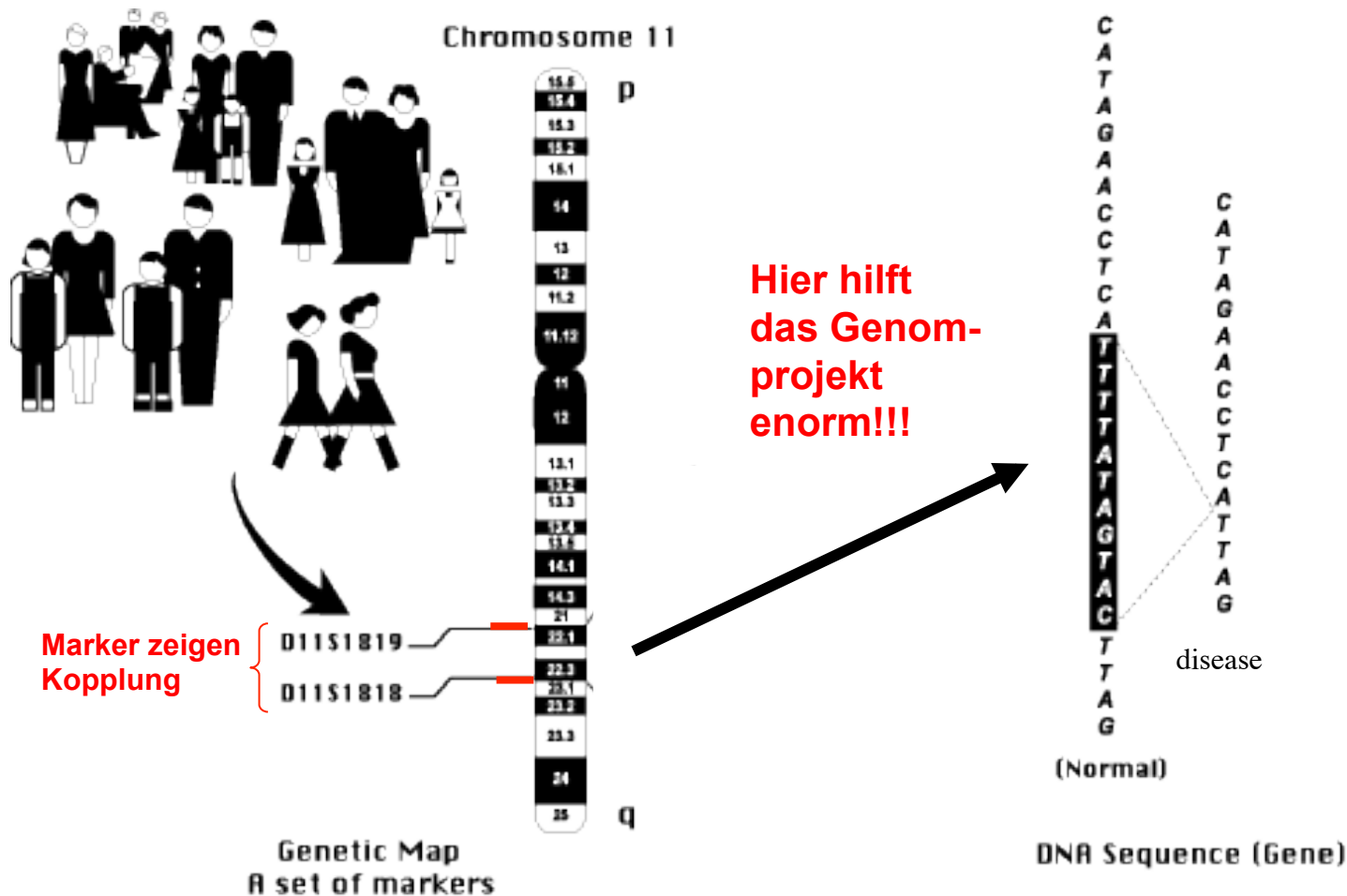
Positionelle Klonierung und Genomsequenzierung ergänzen sich!

Marker-Kopplung zeigt uns ungefähre Lage des Krankheits-Gens an.

Aber wir haben das menschliche Genom doch schon seit 2001 sequenziert!

Q: Wie hilft uns das die Genomsequenz bei der Suche nach dem Krankheitsgen also weiter?

Positionelle Klonierung nutzt die Genomsequenz!



- per Kopplungsanalyse oft Einengung der Genposition auf < 1Mb möglich
- dann Suche nach Mutationen in den „Kandidatengenen“ dieser Region

Speeding up gene discovery

Eiberg et al. 2008

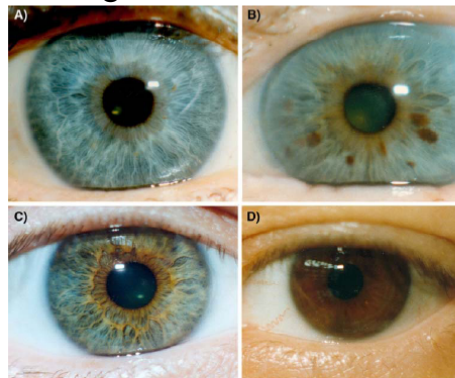


Table 4 Conservation of the eye color silencer sequence in the *HERC2* intron 86 in 9 different species

Species	Eye color	DNA-Library	DNA sequence
Homo	Blue	hg18_dna	TTCATTGAGCAT TAA GTGTCAGTTCTGCACGCTAT
Homo	Brown	hg18_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Chimpanse	Brown	panTro2_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Rhesus monkey	Brown	rheMac2_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Horse	Brown	equCab1_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Cow	Brown	bosTau2_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Cat	Brown-yellow	felCat3_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Dog	Brown-yellow	canFam2_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Rat	Brown	rn4_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Mouse	Brown	Mm8_dna	TTCATTGAGCAT TAA ATGTCAAGTTCTGCACGCTAT
Consensus sequence - blue eye			Ttca-ttg----- Taa GtGtcaa-t-c-----c-tat
Consensus sequence - brown eye			Ttca-ttg----- Taa Atgtcaa-t-c-----c-tat
Nkx-2.5 target site; match allele for blue eye color			TYAAGTG
CdxX-1 target site; match allele for brown eye color			YAKWAWW

The DNA sequences for the 9 species are from the UCSC genome browser (May 2006). The grey shaded sequences represent the binding site region and the bold G or A nucleotide in the gray area represent the variation found between blue and brown eye color. Nkx-2.5 match the blue sequence (score 0.99) and CdxX-1 match brown (score 0.92) and blue sequences (score 0.87) analyzed by TFSEARCH

Family & genetic markers > linkage study > look up candidate gene in genome sequence

Kopplungsanalyse gut wenn...

- Familien/Stammbaum zur Verfügung stehen
- wenige definierte Mutationen den Phänotyp bestimmen
- die „Penetranz“ der Mutation hoch ist

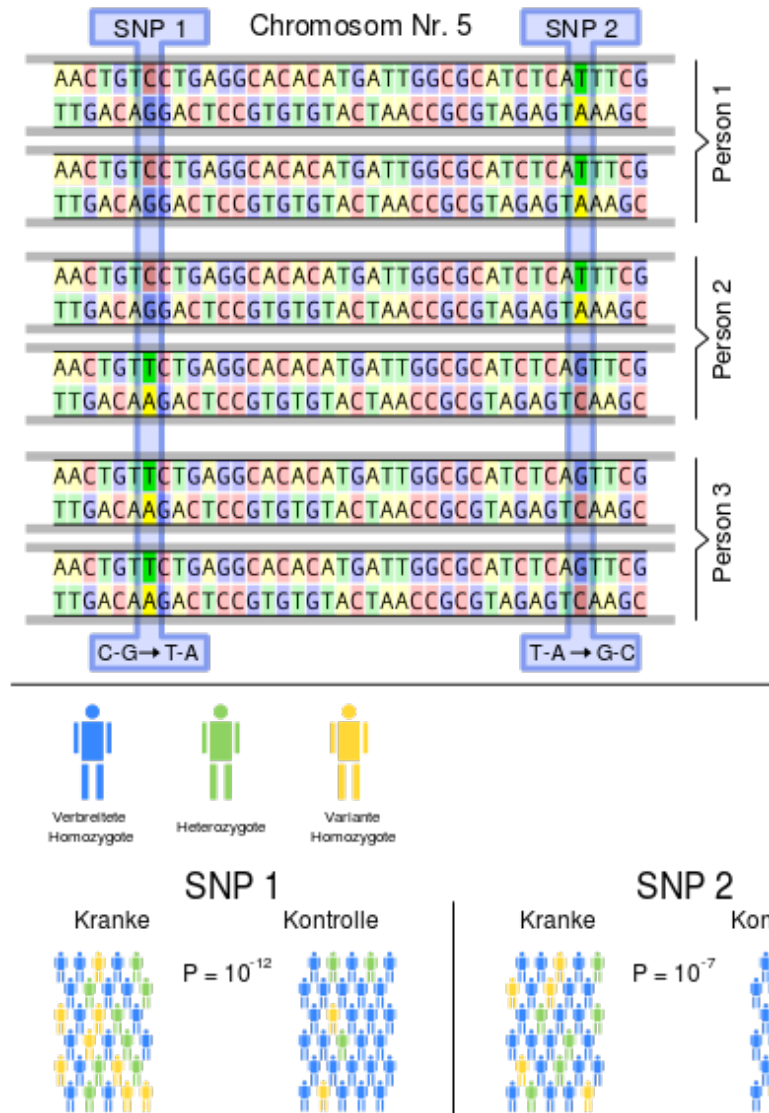
Die meisten verbreiteten genetischen Erkrankungen beruhen aber auf **vielen Mutationen mit geringen Effekten**...(quantitative trait loci)

Genomweite Assoziationsstudien

- Fall-Kontroll-Studie (hunderte bis tausende Probanden)
- sind bestimmte genetische Unterschiede (SNPs) in der Kohorte der Fälle (mit Erkrankung) gegenüber den Kontrollen (Gesunde) gehäuft vorhanden und somit mit der Erkrankung statistisch signifikant assoziiert?
- mit der Erkrankung assoziierte SNPs sind meist nicht direkt für den Phänotyp verantwortlich.

Sie zeigen nur die Lage von relevanten Kandidatengenen, die man sich näher anschauen muss.

Genomweite Assoziationsstudien



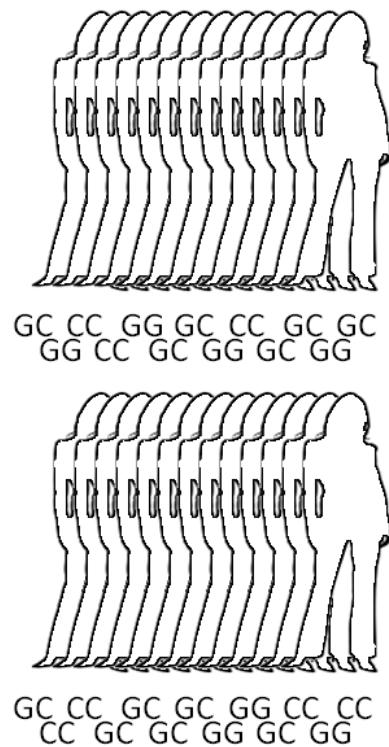
- Hunderte bis Tausende Fälle und Kontrollpersonen > DNA

- Etwa 1 Mio. der ca. 3 Mio. im Genom verteilten SNPs werden in jeder Person sequenziert (SNP-Arrays oder NGS)

- SNP1-Variante kommt in Kranken signifikant gehäuft vor

➤ SNP1 ist mit Krankheits-Gen assoziiert

Genomweite Assoziationsstudien



SNP1

Cases

Count of G:
2104 of 4000

Frequency of G:
52.6%

Controls

Count of G:
2676 of 6000

Frequency of G:
44.6%

P-value:

$5.0 \cdot 10^{-15}$

SNP2

Cases

Count of G:
1648 of 4000

Frequency of G:
41.2%

Controls

Count of G:
2532 of 6000

Frequency of G:
42.2%

P-value:

0.33

SNP...

*Repeat for all
SNPs*

Berechnung
der Signifikanz

Chi-Quadrat Test

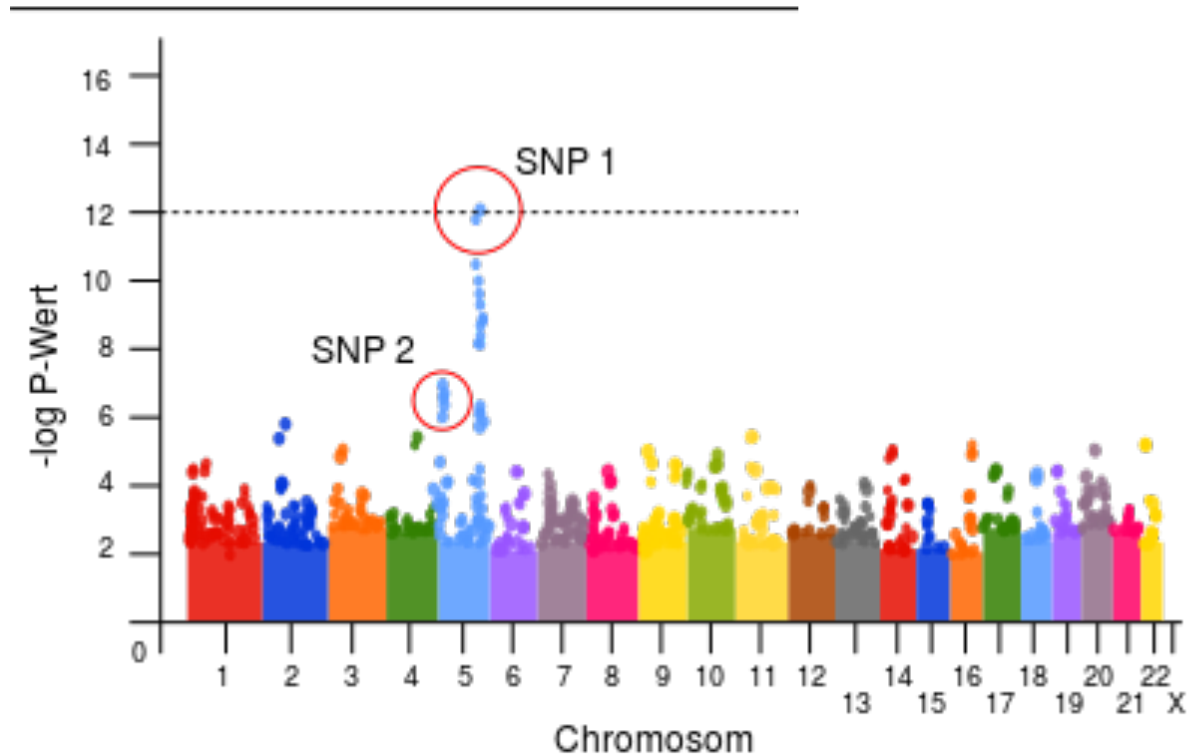
Beispiel aus Wikipedia nach Daten aus:

Wellcome Trust Case Control Consortium, Burton PR (June 2007)

[Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls".](#)

[Nature. 447 \(7145\): 661–78.](#)

Genomweite Assoziationsstudien



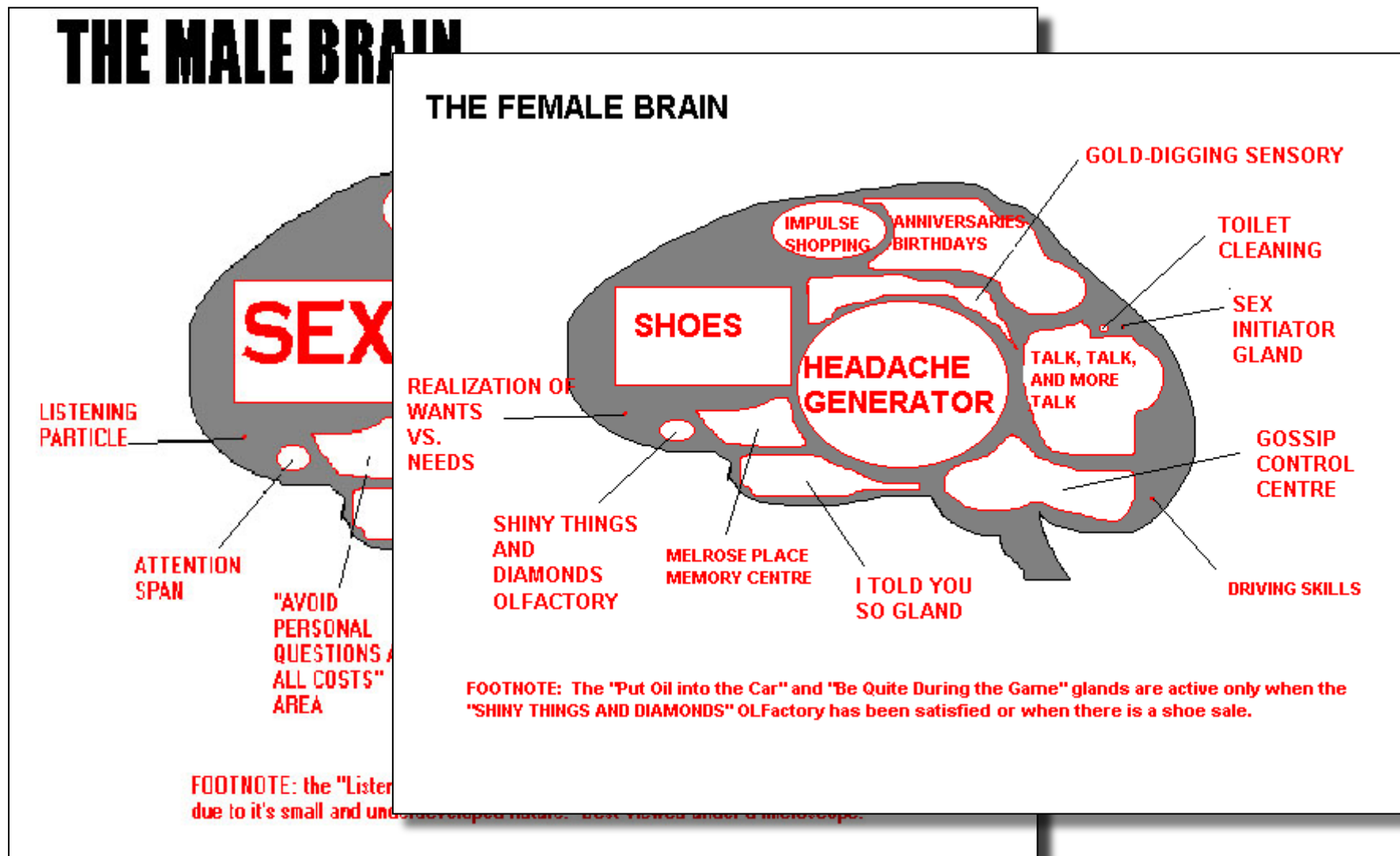
„Manhattan-Plot“ zeigt Position von SNPs, die signifikant mit dem Phänotyp assoziiert sind

Wege zu „allen Genen“ durch Genomanalyse (Genomik)

- **Gesamtgenom-Sequenzierung**
„whole genome shotgun“
- **EST-Sequenzierung**
 - > „**e**xpressed **s**equences **t**ags“, cDNA- ‘Schnipsel ‘
 - > Gen-Entschlüsselung ,leicht gemacht ‘
 - > genannt **RNA-Seq** bei Verwendung von NGS-Technik

Diese beiden Ansätze komplementieren einander...

Welche Gene sind da aktiv...?



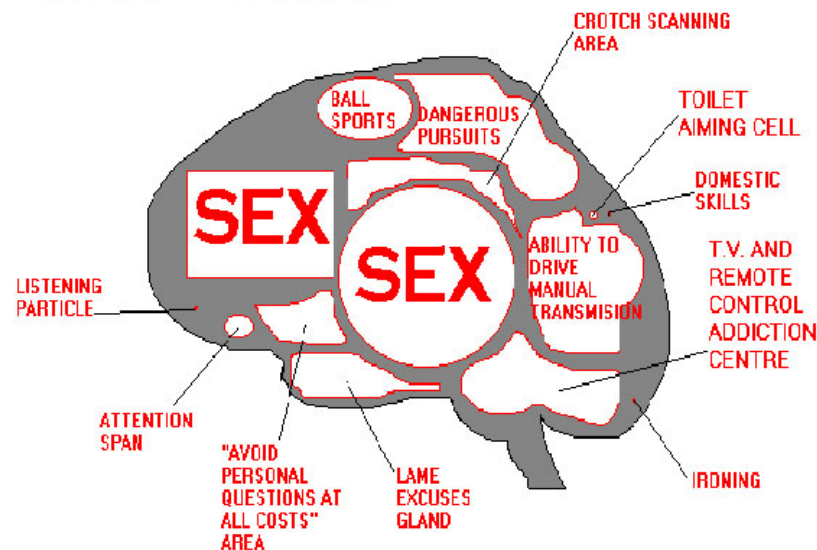


J. C. Venter

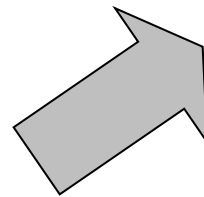
EST-Sequenzierung

ESTs sind „quick-and-dirty“ sequenzierte cDNA-Schnipsel

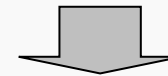
THE MALE BRAIN



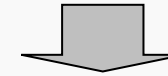
FOOTNOTE: the "Listening to children cry in the middle of the night" gland is not shown due to its small and underdeveloped nature. Best viewed under a microscope.



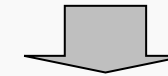
Gewebe (z.B. Gehirn)



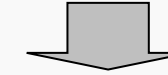
RNA-Präparation



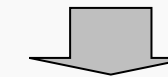
Umschreiben in cDNA



(cDNA-Klonbank)

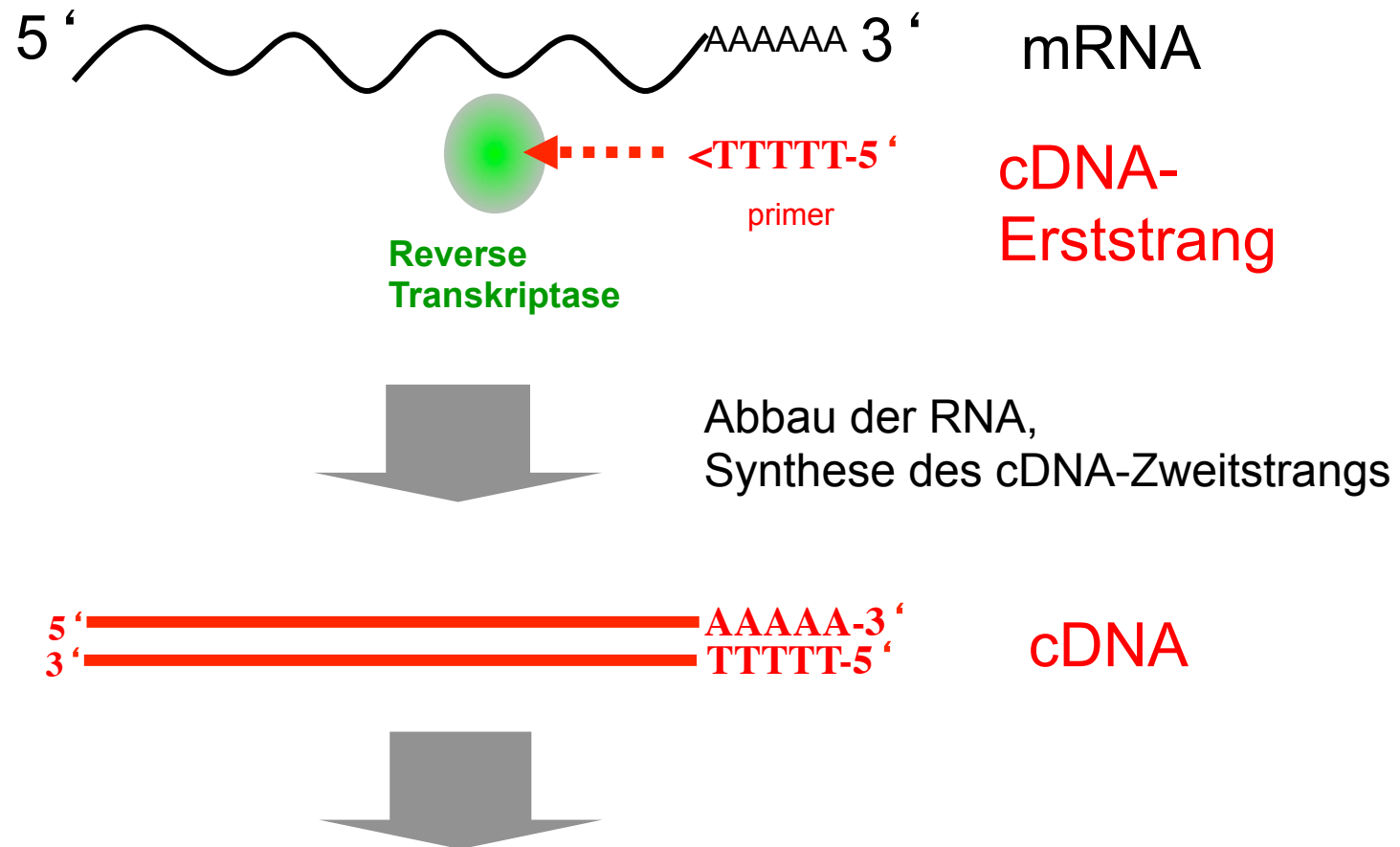


wahlloses Sequenzieren
der cDNAs



Genkatalog

Herstellung einer cDNA-Bibliothek



traditionell: Klonierung in (Plasmid)-Vektoren > **EST-Seq**
heute (NGS): direkte Sequenzierung ohne Klonierung (**RNA-Seq**)

Gen, cDNA, EST

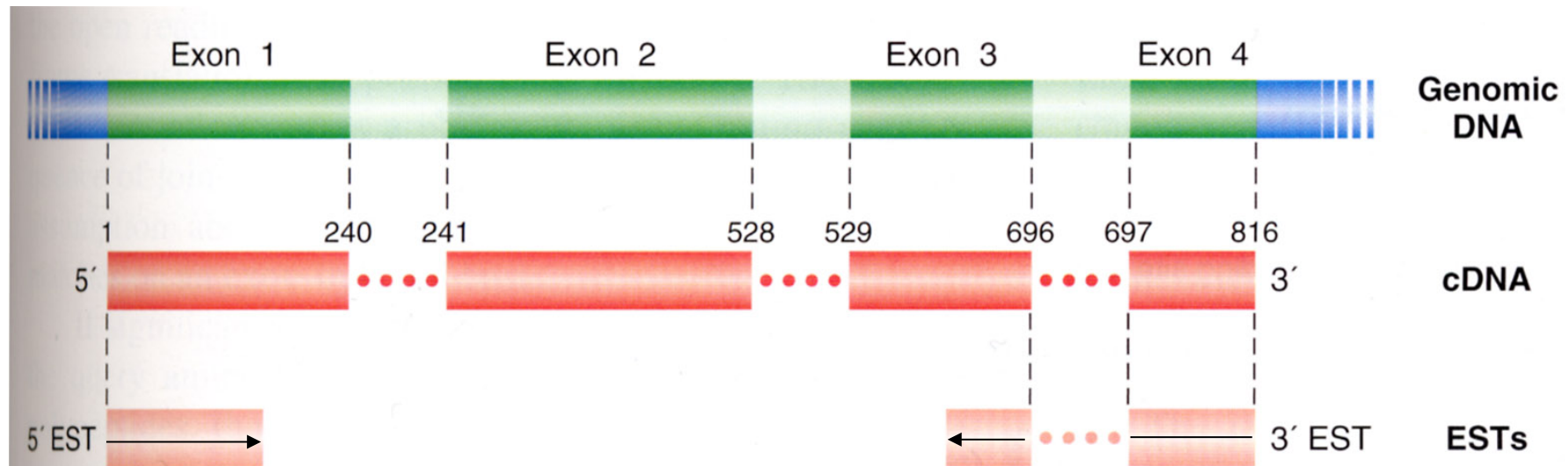


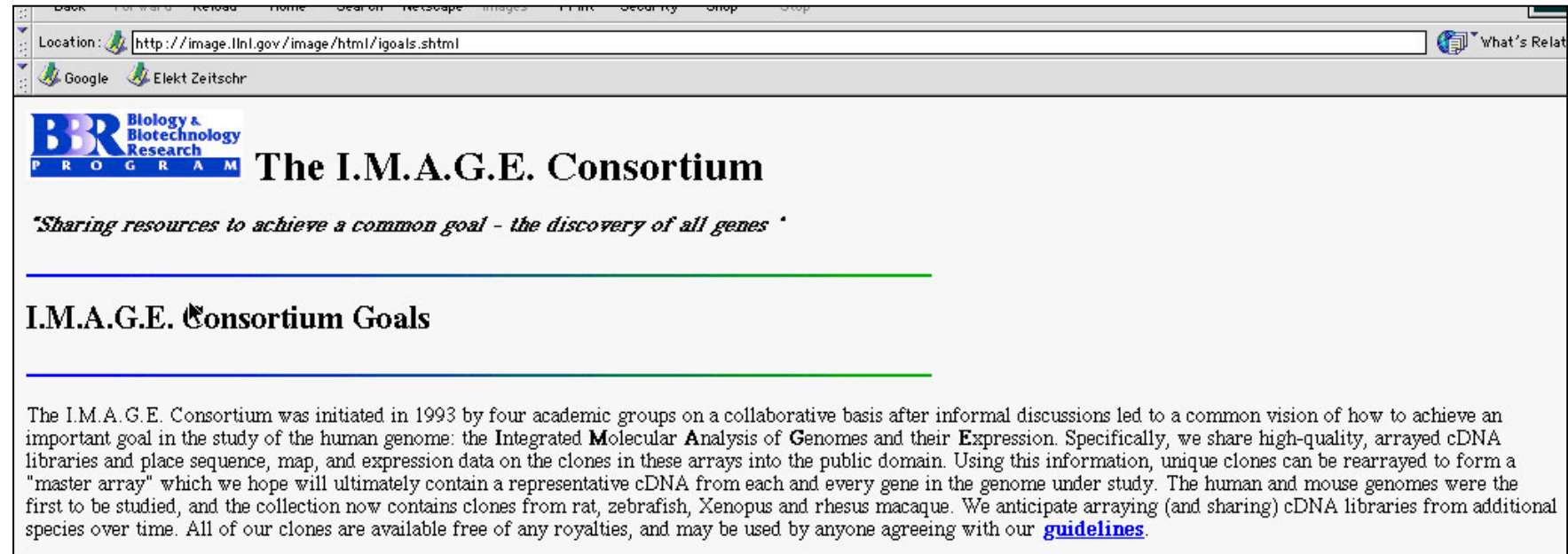
Figure 9-32 The alignment of fully sequenced cDNAs and ESTs with genomic DNA. The solid lines indicate regions of alignment; for the cDNA, these are the exons of the gene. The dots between segments of cDNA or ESTs indicate regions in the genomic DNA that do not align with cDNA or EST sequences; these are the locations of the introns. The numbers above the cDNA line indicate the base coordinates of the cDNA sequence, where base 1 is the 5'-most base and base 816 is the 3'-most base of the cDNA. For the ESTs, only a short sequence read from either the 5' or 3' end of the corresponding cDNA is obtained. This establishes the boundaries of the transcription unit, but it is not informative about the internal structure of the transcript unless the EST sequences cross an intron (as is true for the 3' EST depicted here).

→ Sequenz-Read

Gen, cDNA, EST

Q: warum erscheint es wenig attraktiv, eine 3'-Sequenzierung von EST-Klonen durchzuführen?

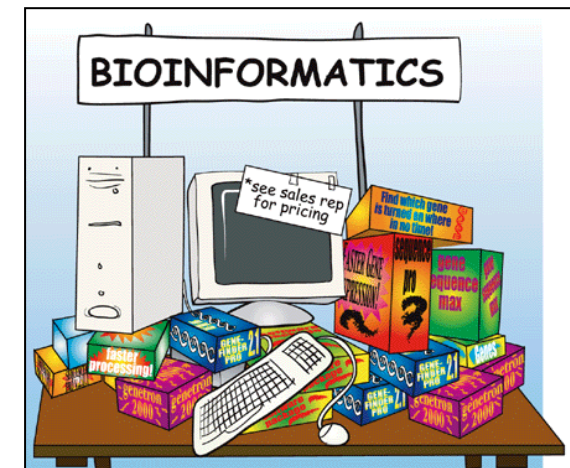
EST-Projekte



- unzählige EST-Projekte für diverse Arten und diverse Organe/ Entwicklungsstadien/ Krebs etc...
- **EST-Klone sind öffentlich erhältlich!** www.imageconsortium.org
- RNA-Seq-Daten (NGS) finden sich weiteren, speziellen Datenbanken (SRA, GEO)

Warum Datenbanken?

- Sammeln und Erhalten von Daten
- Daten such- und auffindbar machen
- Daten-Darstellungsweise vereinheitlichen
- **aus Daten Wissen machen!**



Datenbanken in der Molekularbiologie

- Literaturdatenbanken
- Sequenzdatenbanken
 - primäre DB: annotierte DNA- und Proteinsequenzen
 - abgeleitete DB: interpretierte Sequenzdaten
(z.B. Proteindomänen oder Stoffwechselwege)

Ask an Expert

Post your genetics question

Track your questions **POST**

Explore

Within this Topic (26)

- Comparative Genomics (5)
- Genome Sequencing and Annotation (6)
- Functional Genomics (4)
- Translational Genomics (6)

Or **Browse Visually**

Related Topics

Genetics

- Chromosomes and Cytogenetics
- Evolutionary Genetics
- Gene Expression and Regulation

READING

GENOMICS | Lead Editor: Michael Goldman, Christopher D. Smith



Genomic Data Resources: Challenges and Promises

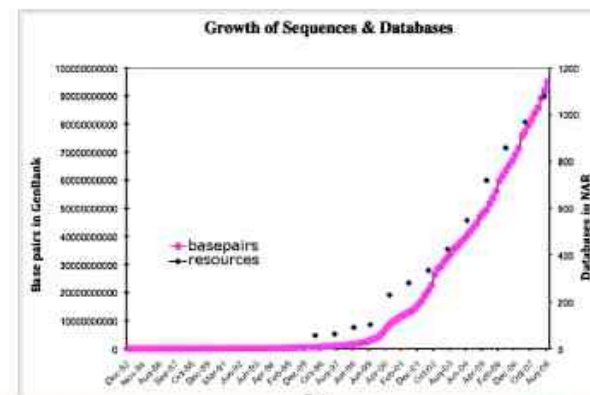
By: Warren C. Lathe III (*OpenHelix*), Jennifer M. Williams (*OpenHelix*), Mary E. Mangan (*OpenHelix*) & Donna Karolchik (*University of California, Santa Cruz Genome Bioinformatics Group*) © 2008 Nature Education

Citation: Lathe, W., Williams, J., Mangan, M. & Karolchik, D. (2008) Genomic Data Resources: Challenges and Promises. *Nature Education* 1(3)

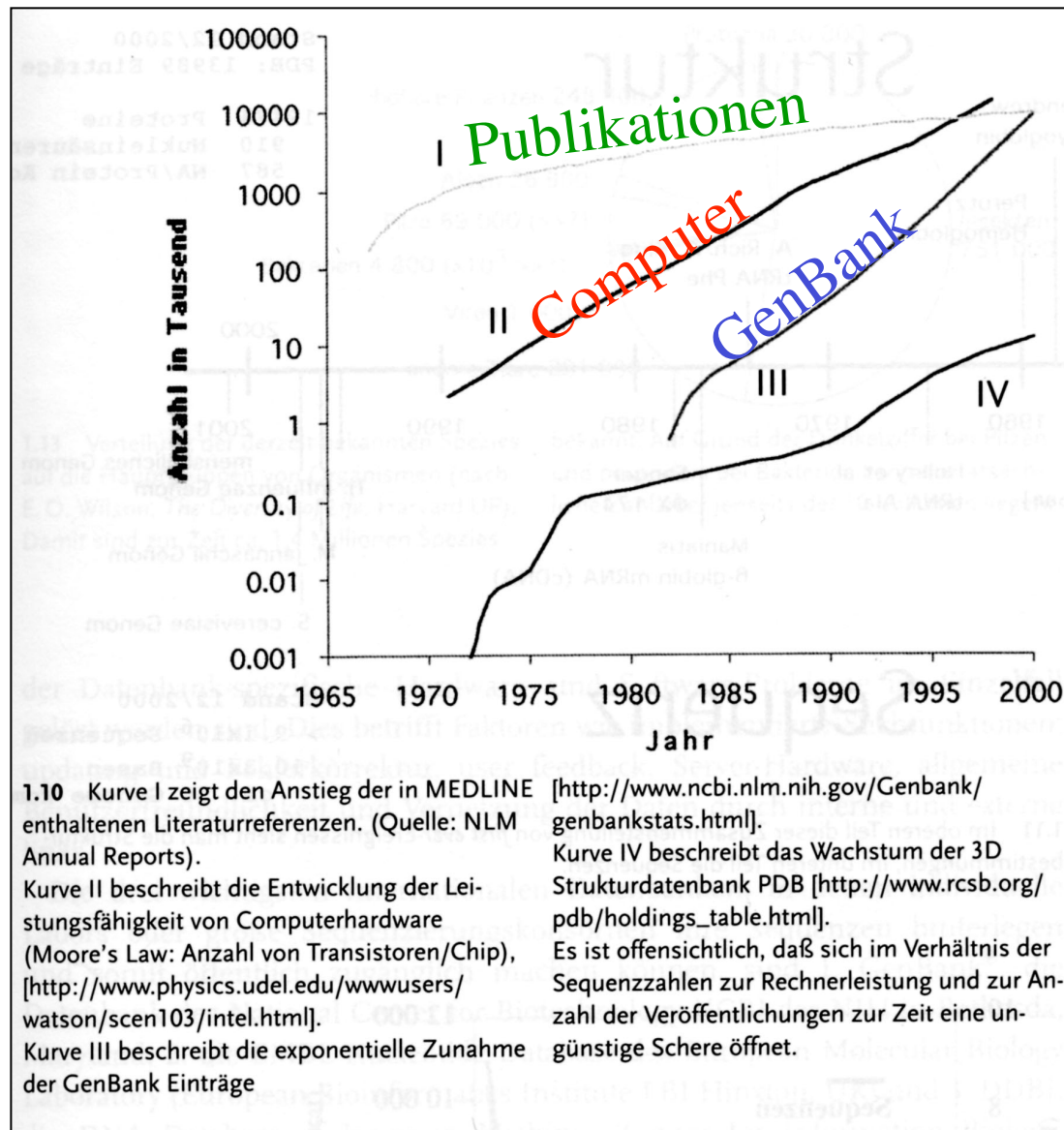
Where would you go to find the nucleotide and amino acid data you need? There are thousands of genomic databases, tools, and other resources freely accessible on the Internet.

Computer databases are an increasingly necessary tool for organizing the vast amounts of biological data currently available and for making it easier for researchers to locate relevant information. In 1979, the Los Alamos Sequence Database was established as a repository for biological sequences. In 1982, this database was renamed GenBank and, later the same year, moved to the newly instituted **National Center for Biotechnology Information (NCBI)**, where it lives today. By the end of 1983, more than 2,000 sequences were stored in GenBank, with a total of just under 1 million **base pairs** (Cooper & Patterson, 2008).

At about the same time, a joint effort between NCBI, the **European Molecular Biology Laboratory (EMBL)**, and the **DNA Databank of Japan (DDBJ)** created the **International Nucleotide Sequence Database Collaboration (INSDC)** to collect and disseminate the burgeoning

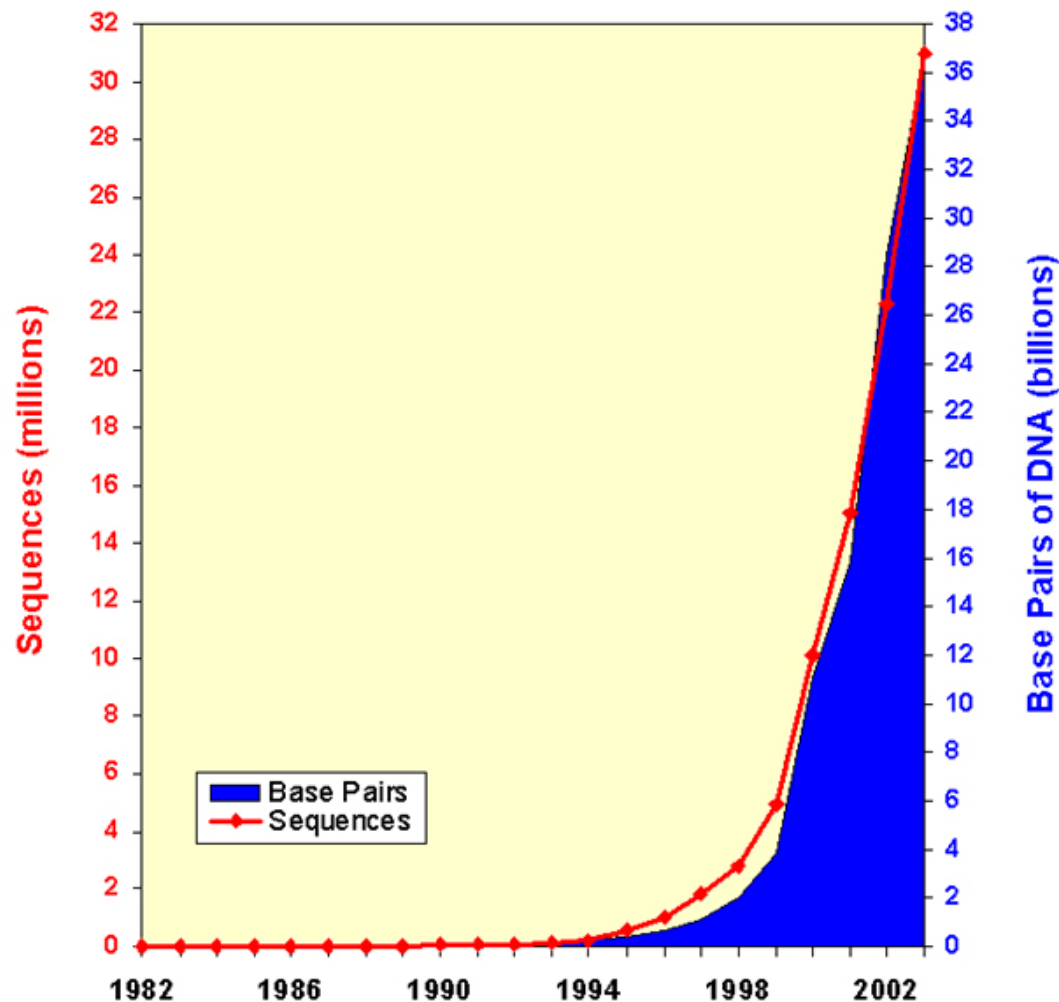


Datenbanken- Wachstum



Datenbank-Wachstum

Growth of GenBank



22,617,000,000 bases in
18,197,000 sequence records
(August 2002)

145,500,000,000 bases in
158,000,000 sequence records
(Oct 2012)

Bitte Stand Okt 2018
selbst recherchieren!

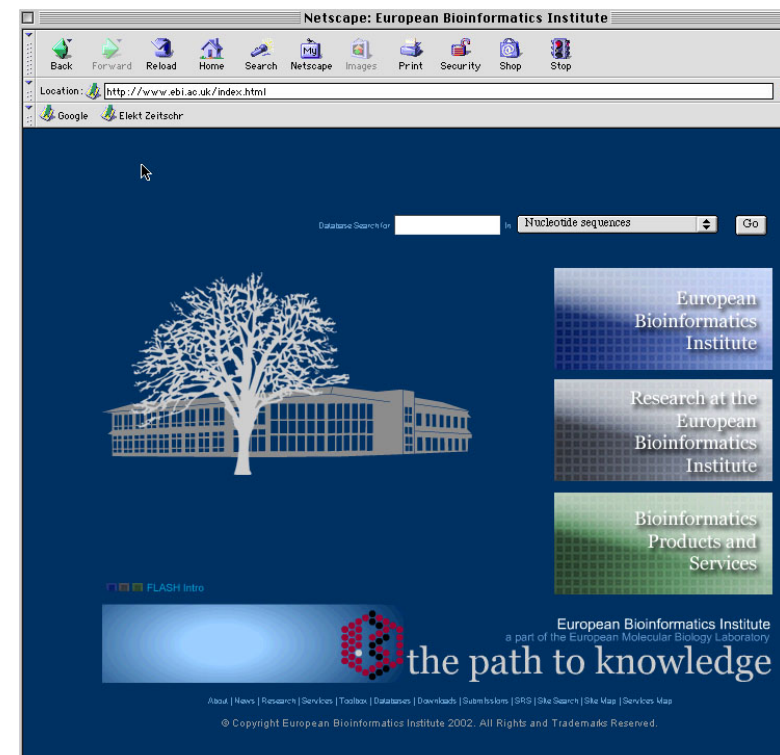
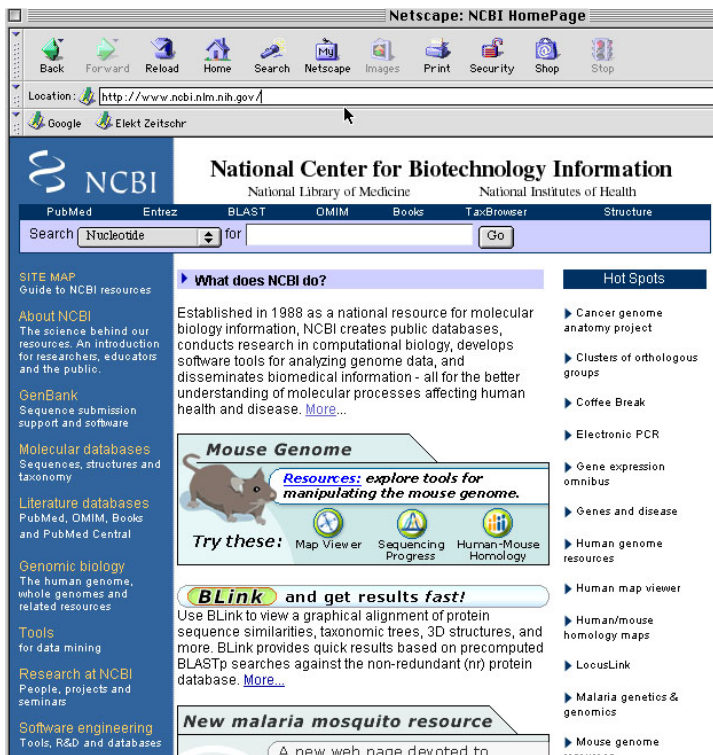
Datenbanken in der Molekularbiologie

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

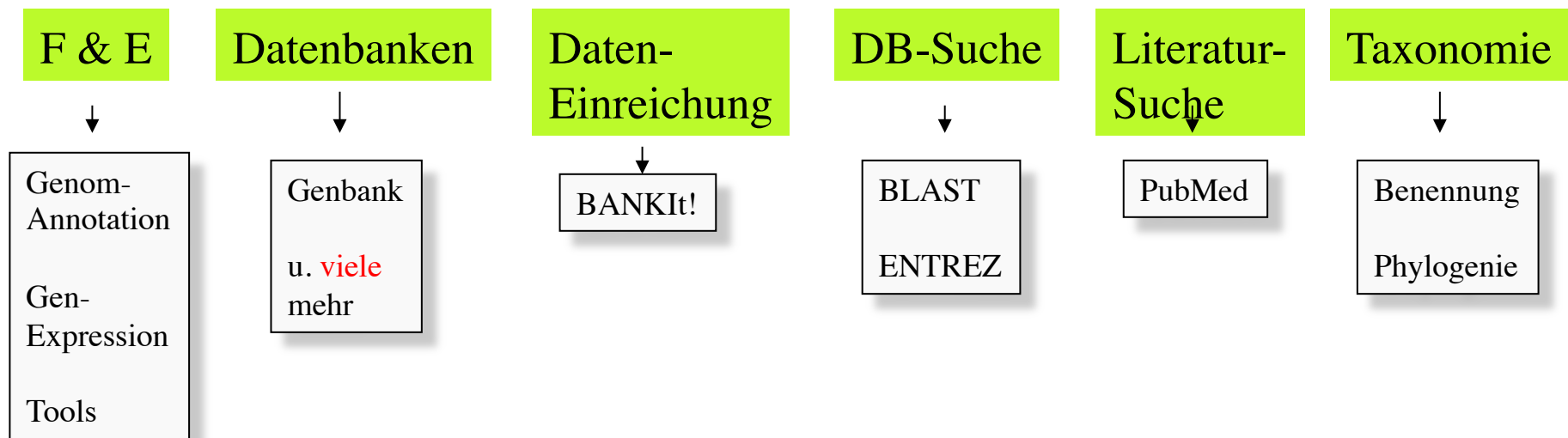
National Center for Biotechnology Information,
Am NIH, Bethesda, Maryland, USA

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

European Bioinformatics Institute,
Sanger Campus, Hinxton, GB



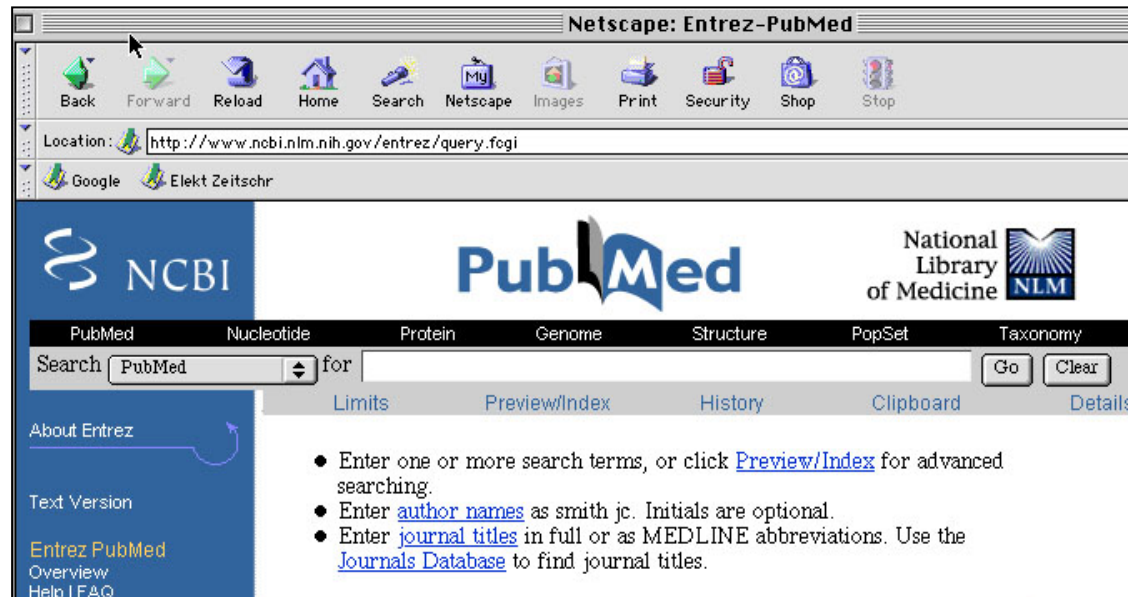
Leistungen des NCBI im Überblick



...sowie Trainings, Tutorials und vieles mehr“!!

Literatur-Datenbank und -Suche

- PubMed = Public Medline /NCBI



- Suchdienst der Natl. Library of Medicine
- > 5600 biomedizinische Zeitschriften
- Verbindung zu Online-Zeitschriften > Download!! **VPN starten!**
- Suchbegriffe einfach eingeben
(Boole'sche Verknüpfungen: „AND“, „OR“, „NOT“; Truncation: „*“)

Deja vu

A study of scientific publication
ethics

[Home](#)[Browse](#)[Report](#)[Help](#)[Statistics](#)[Contact Us](#)[Team](#)[References](#)

Powered by eTBLAST

Innovation Labs

UT Southwestern Medical School - Dallas

62 000 Arbeiten durchsucht:
0.04% Plagiate
1.35 % Duplikate

Deja Vu: a Database of Highly Similar Citations*

Click [this link](#) to begin browsing entries, or click the "Browse" button above and follow the instructions.

We value your feedback. Please take one minute to take a brief survey (Click [here](#)). We appreciate your support.

Deja vu is a database of extremely similar **Medline** citations. Many, but not all, of which contain instances of duplicate publication and potential plagiarism. Deja vu is a dynamic resource for the community, with manual curation ongoing continuously, and we welcome input and comments.

Latest News

**2009-05-27 - Nature
Medicine News**

Deja vu is in Nature Medicine News. [Read it.](#)

**2009-05-22 - On Science
NEWS FOCUS**

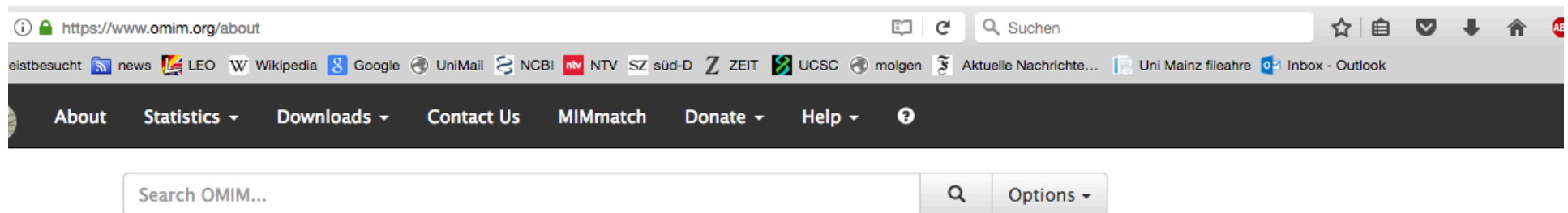
A large article "Plagiarism Sleuths" published in Science News Focus brings up Deja vu's work and broad discussions. [Read it.](#)

**2009-03-06 - Deja Vu in
Science**

Our article "Responding to Possible Plagiarism" was published

OMIM: eine spezielle Literatur-Datenbank

Online Mendelian Inheritance of Man



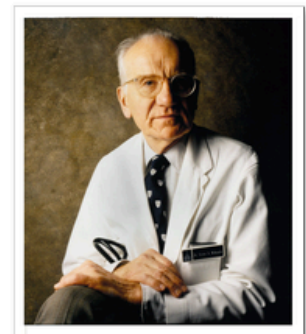
OMIM® - Online Mendelian Inheritance in Man®

Welcome to OMIM®, Online Mendelian Inheritance in Man®. OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 15,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.

OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh.

[NLM's Profiles in Science -- The McKusick Papers](#)



OMIM: eine spezielle Literatur-Datenbank

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

OMIM Entry Statistics

Number of Entries in OMIM (Updated October 23rd, 2017) :

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	14,946	725	49	35	15,755
Gene and phenotype, combined +	76	0	0	2	78
Phenotype description, molecular basis known #	4,753	320	4	31	5,108
Phenotype description or locus, molecular basis unknown %	1,467	124	5	0	1,596
Other, mainly phenotypes with suspected mendelian basis	1,665	107	2	0	1,774
Totals	22,907	1,276	60	68	24,311

Integrierte Such-Werkzeuge!

NCBI Resources ▾ How To ▾ Sign in to NCBI

NCBI will begin redirecting all HTTP traffic to HTTPS on Thursday, November 10 at 9 AM EST (2 PM UTC). [Read more.](#)

Search NCBI databases Help

Search

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

probieren Sie mal: SARS

www.ncbi.nlm.nih.gov/Entrez/

...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 rllvvypwtg rffesfgdls tpdavmgnpk
agtcctttgg ggatctgtcc actcctgatg ctgttatggg caaccctaag gtgaaggctc

...kommt erst in 2 Wochen!

Sequenz-Datenbanken

- komplette Übersicht: Januar-Ausgabe von Nucleic Acids Research



z. B.

„Genbank“

„Flybase“

„Wanda“: a library of duplicated fish genes“

„ENZYME“

- www.oxfordjournals.org/nar/database/a/
- > 1500 Datenbanken in der Sammlung!!

Sequenz-Datenbanken

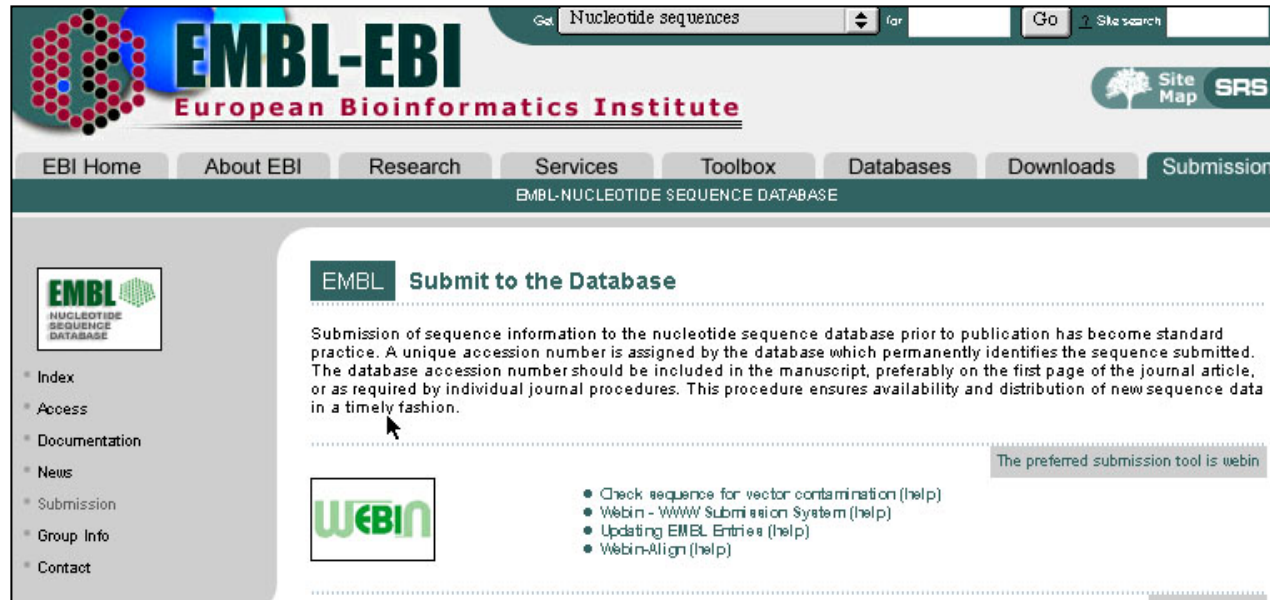
NCBI > GenBank (1979)

EBI > EMBL database (1980)
heute: 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

Eintragen in Sequenz-Datenbanken



WEBIN


[www.ebi.
ac.uk/submissions](http://www.ebi.ac.uk/submissions)

- Jeder darf seine Sequenzen unbeschränkt (und leider oft auch unkorrigiert) eintragen!
(> Kontaminationen der DB mit Vektorsequenzen u. ä.)
- Korrekturen müssen bei Ort des Eintrags durchgeführt werden


Für neue menschliche Gene: Vor dem Eintrag den Namen prüfen und schützen lassen!

http://www.gene.ucl.ac.uk/nomenclature/

Elekt Zeitschr




HUGO Gene Nomenclature Committee



[Home](#) | [About HGNC](#) | [Database](#) | [Guidelines](#) | [Submissions](#) | [Downloads](#) | [Gene Families](#)

Giving unique and meaningful names to every human gene


[Commercial Users](#)
[Contact Us](#)
[Database Links](#)
[FAQs](#)
[IAC](#) (International Advisory Committee)




Search Approved Symbols

We have approved symbols for nearly one half of the genes in the human genome and, with an estimated 15,000 more genes to name, we still have plenty to do! Use the Genew database to search for your gene.

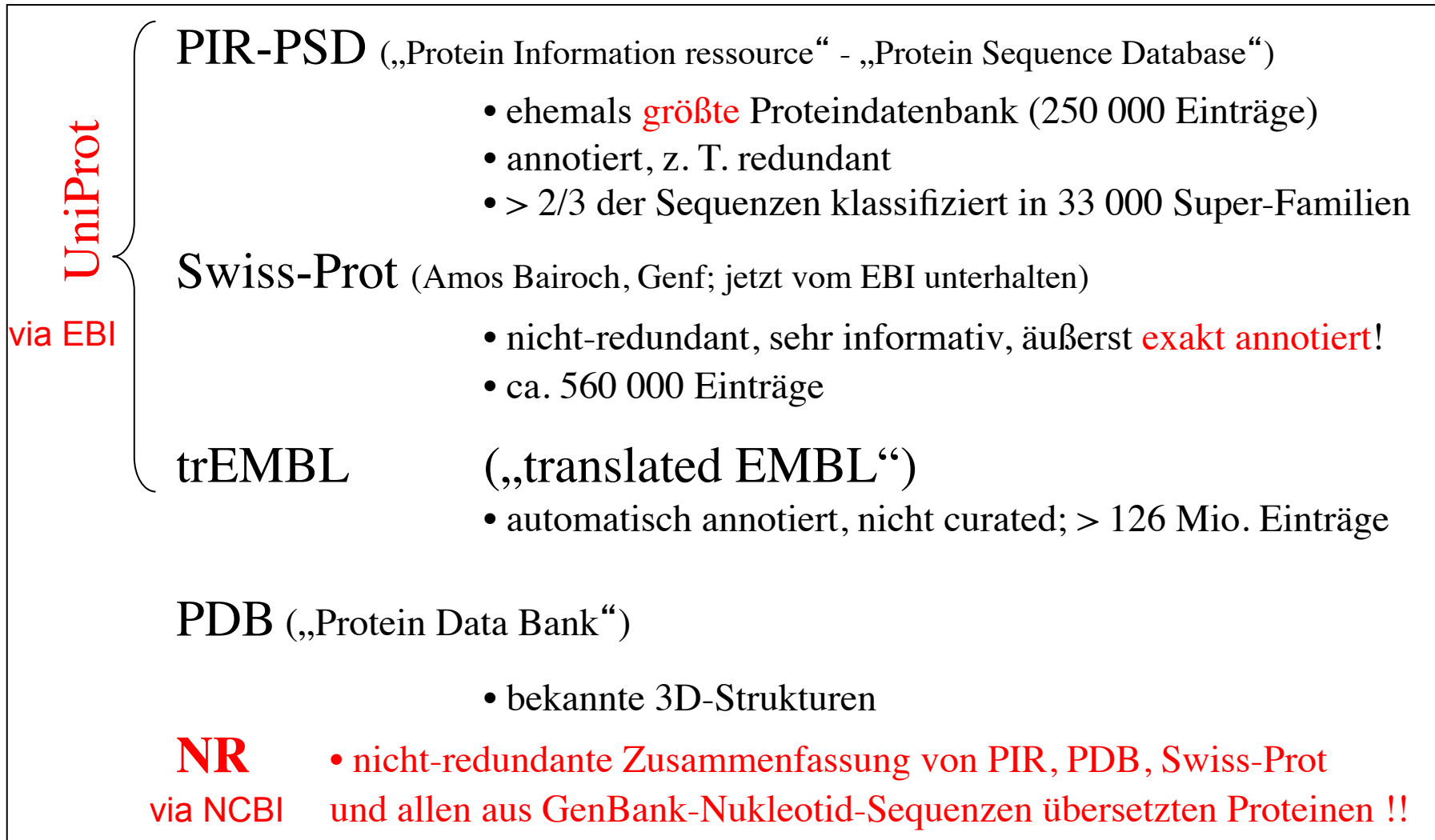
Quick Gene Search



[BLAST](#)
[Entrez](#)
[LocusLink](#)
[OMIM](#)



Protein-Sequenzdatenbanken



DNA-Sequenzdatenbanken

(via NCBI)

GenBank

- > 260 Milliarden Nukleotide
- Größe verdoppelt sich alle 10-12 Monate
- > 1000 komplett sequenzierte Bakterien-Genome
- ca 500 annotierte Eukaryotengenome (in Arbeit)
- 67% der eingetragenen Sequenzen sind ESTs (1200 Spezies)
- > 300 000 Spezies repräsentiert (2200 neue/Monat)
- „top organism“: Mensch mit 16 Milliarden Bp
- 30 000 Zugriffe pro Tag
- GenBank ist in 20 Abteilungen unterteilt !

GenBank-Unterteilungen

NR
(Nucleotide)

- ~~nicht-redundante~~ Zusammenfassung aus GenBank + EMBL + DDBJ + PDB
- die Abteilungen EST/HTGS etc sind **AUSGESCHLOSSEN!!**

dbEST • redundante Datenbank aller beim NCBI, EBI und in Japan eingereichter EST cDNA-Sequenzen

dbSNP • Datenbank für SNP-Marker

HTG • „high throughput genomic sequences“ aus Genomprojekten; nur Sanger-Reads

SRA • sequence read archive: nur NGS-Rohdaten

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

sowie ALU (repeats), VECTOR, YEAST, MITO, PAT(ente) und mehr

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 191 (8/2012)	Annual increase (%) ^a
Taxnomic divisions			
SYN	Synthetic	928 200 038	494.2%
PHG	Phages	84 079 451	34.4%
ENV	Environmental samples	3 374 433 548	32.1%
VRL	Viruses	1 429 464 786	21.1%
BCT	Bacteria	8 439 854 434	21.0%
PLN	Plants	5 481 470 133	15.6%
MAM	Other mammals	863 036 872	6.9%
VRT	Other vertebrates	2 886 594 595	6.7%
PRI	Primates	6 317 656 773	3.3%
UNA	Unannotated	127 803	1.5%
ROD	Rodents	4 435 106 948	0.9%
INV	Invertebrates	2 493 058 927	-1.7%
Functional divisions			
TSA	Transcriptome shotgun data	5 759 588 580	207.3%
WGS	Whole-genome shotgun data	308 196 411 905	47.9%
PAT	Patented sequences	12 118 622 726	8.6%
GSS	Genome survey sequences	21 947 780 105	5.7%
EST	Expressed sequence tags	40 888 051 100	4.8%
HTG	High-throughput genomic	24 359 210 558	0.1%
STS	Sequence tagged sites	636 262 446	0.1%
HTC	High-throughput cDNA	639 165 410	-3.5%
TOTAL	All GenBank sequences	451 278 177 138	33.1%

^aMeasured relative to Release 185 (8/2011).

Ein GenBank-Flat File...

accession no.

Version

GI-Nr. ist singulär!

Zitat

CDS = coding sequence

Protein

```

1: AJ315164. Mus musculus Cygb.
LOCUS       MMU315164               9488 bp    DNA     linear   ROD 09-JUL-2002
DEFINITION  Mus musculus Cygb gene for cytoglobin.
ACCESSION   AJ315164
VERSION     AJ315164.1  GI:21727817
KEYWORDS    CYGB gene; cytoglobin.
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1
  AUTHORS   Ebner,B., Burmester,T. and Hankeln,T.
  TITLE     Comparative sequence analysis of the mouse cytoglobin gene
  JOURNAL   Unpublished
REFERENCE   2 (bases 1 to 9488)
  AUTHORS   Hankeln,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-JUL-2001) Hankeln T., Inst. Molekulargenet., Univ.
            Mainz, J.J. Becherweg 32, Mainz, D-55099, GERMANY

FEATURES             Location/Qualifiers
     source            1..9488
                       /organism="Mus musculus"
                       /db_xref="taxon:10090"
     gene              1736..8693
                       /gene="Cygb"
     mRNA              join(<1736..1878,5637..5868,6206..6369,8660..>8693)
                       /gene="Cygb"
     CDS               join(1736..1878,5637..5868,6206..6369,8660..8693)
                       /gene="Cygb"
                       /codon_start=1
                       /product="cytoglobin"
                       /protein_id="CAC86190.1"
                       /db_xref="GI:21727818"
                       /translation="MEKVPGDMEIERRERSEELSEAERKAVQATWARLYANCEDYGV
ILVRFVNFPSAKQYFSQFRHMDPLEMERSPOLRKHACRVMGALNTVVENLHDPDKV
SSVLALVGKAHALKHKEVPMYFKILSGVILEVIAEEFANDFPVETQKAWAKLRGLIYS
HVTAAAYKEVGWVQVNPNTITPPATLPSSGP"
     exon              <1736..1878
                       /gene="Cygb"
                       /number=1
     intron            1879..5636
                       /gene="Cygb"
                       /number=1
     exon              5637..5868
                       /gene="Cygb"
                       /number=2
     intron            5869..6205
                       /gene="Cygb"
     intron            6206..6369
                       /gene="Cygb"
                       /number=3
     intron            6370..8659
                       /gene="Cygb"
                       /number=3
     exon              8660..>8693
                       /gene="Cygb"
                       /number=4

BASE COUNT      2078 a      2830 c      2633 g      1947 t
ORIGIN
1  ttttgttatt agtgtgtgtg tgtgtgtgtg tgtgtgtgtg tgtgtgtgtg tgtgtgtgtg
61  tgtgagaaag gacagcttgt aagagtcaat ctcggtctgg tgagatggca cagtgggtaa
121  gagcaccoga ctgctcttct gaaggtccgg agttcaaatc ccagcaacca catggtggct
181  cacaaccatc cgtaaacaaga tctgacgccc tcttctggag tgtctgaaga cagctacagt
241  gtacttcatc ataaataaat aaatctttaa aaaaaaaaaa aaaaagagtc aatctcttcc
301  ttccaccocg tgggtctctag ggcgggaact cttcagatca tcaagttttg tgaggcaaat
361  acctttaaag attttaaatc aaacagggac caagctgaag aggagagcca ctacttcct
421  gggcagcagg gctcctcttc atcttgagct gggagccttg aggagaacca gagacagtgc
481  acttatctgt gtcaggacca ggcaggccct ctgtgtcttc agggctccct ctgtctacag
541  ccaggtctga gcaactgctg gcagggtgag ggttctgggt ctctcaagac tgcactcctc
601  cctctgtgcc cagtgctcac tctccttgag cagtctaaga aggagatgaa ggatctgcct
661  tctgtgttct aaactgaact ctcgatgggt acaaatgtct tcactgtccg gtgccttact
721  caggacttcc ggctccccag agcctctcca tcatacctga ctgactgcct ctctgtggaa
781  ctctactcca ccacggctag ggctaggacg gtaattcagg acagtgtctg gtccttatct
841  acctatcatt aactaccttc tcaggactcc ctgctcgagc gcgtaggagg ggaagtgggg
901  caggggtctc ttggtcccca gactggacat gtgcacaagt cacaagagc cagctggacc
961  agcagatggg aagagagact gccactgcct ctaaaacctg cagaactcga ggtccccacg
1021  tctgacaaag ggtgctctcg ccagctctca ggctcaggag tggcggggct acaggggaag
1081  ccggatttgg tctgcaagtt cctcctccca agcaggcgat gcctgggtgt gtttgcattt
1141  gagtacagtc ccaactcgtg gaagggtgta cacacacaca cacaacaca cagctccttc
1201  cacacacgtc cagacaccac acgtcgagag ctgagacccg caccctcagc tccgcacag
1261  ccggggggcg accgcagtag acccgcgctc gtcctacacc gcgtgacccc ctgatcctcc
1321  cagccctctc tgcacattgg ccacacctac ctctccagc cgggacccgg gtggccttgc
1381  taacgggtgg gtggtcaggc aggcagacgc cagcogtgac accccactcc cgcctaacct
1441  tcaccttgcc aaaaattgact ccagaaaaa cgaactggatt ttttgagcg gatthttttt
1501  aaaaaacatt ttttccccag cagacacatc ctccgcccc cagctcgagc ccccgcccc
1561  ccgcacatat accctgcaca ccgcgcgcga cacacacccg gcgcgcacag acacacgctc
1621  cctcctctcg gcgtctact cctcgccgcg ccgcctctct gcccgctcgc tgcctcagga
1681  cctcggtctc ccgcgcgcgc gcgcagcgca agctcgcgct ggctcgagcg cagctcagga
1741  gaaagtccgc ggcgcacatg agatagagcg tagggagagg agcagaggag tgtccgaggc
1801  ggagaggaag gcggttcagg ctacgtgggc ccgctctgat gccacctcgc aggcctgggc

```

...die dazugehörige Proteinsequenz

Back Forward Reload Home Search Netscape Images Print Security Shop Stop

Location: <http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=10864065&db=Protein&dopt=GenPept>

Google Elekt Zeitschr

☐ 1: NP_067080. neuroglobin [Homo...[gi:10864065]

LOCUS NGB 151 aa linear PRI 05-NOV-2002
DEFINITION neuroglobin [Homo sapiens].
ACCESSION NP_067080
VERSION NP_067080.1 GI:10864065
DBSOURCE REFSEQ: accession [NM_021257.2](#)
KEYWORDS
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (residues 1 to 151)
AUTHORS Burmester, T., Weich, B., Reinhardt, S. and Hankeln, T.
TITLE A vertebrate globin expressed in the brain
JOURNAL Nature 407 (6803), 520-523 (2000)
MEDLINE [20479975](#)
PUBMED [11029004](#)
COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final
NCBI review. The reference sequence was derived from [AF422797.1](#).
FEATURES
source Location/Qualifiers
1..151
/organism="Homo sapiens"
/db_xref="taxon:9606"
/chromosome="14"
/map="14q24"
[Protein](#) 1..151
/product="neuroglobin"
[Region](#) 7..147
/region_name="Globin"
/note="globin"
/db_xref="CDD:[pfam00042](#)"
[CDS](#) 1..151
/gene="NGB"
/coded_by="NM_021257.2:376..831"
/db_xref="LocusID:[58157](#)"
/db_xref="MIM:[605304](#)"
ORIGIN
1 merpepelir qswravsrsp lehgtvlfar lfalepdllp lfgyncrqfs spedclsspe
61 fldhirkvml vidaavtnve dlssleeyla slgrkhravg vklsfstvg esllymlekc
121 lgpafptr aawsqlygav vqamsrgwdg e
//

Proteinfamilie

...und das Ganze als „FASTA“-Flatfile-Format

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

```
>gi|11967939|ref|NP_071859.1| neuroglobin [Mus musculus]
MERPESELIHQSWRVVSRSPLEHGTVLFARLFALPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLTSLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDPFTPATRTAWSRLYGAV
VQAMSRGWDGE
```

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

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

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

dbEST release 130101

Summary by Organism - 01 January 2013

Number of public entries: 74,186,692

Homo sapiens (human)	8,704,790
Mus musculus + domesticus (mouse)	4,853,570
Zea mays (maize)	2,019,137
Sus scrofa (pig)	1,669,337
Bos taurus (cattle)	1,559,495
Arabidopsis thaliana (thale cress)	1,529,700
Danio rerio (zebrafish)	1,488,275
Glycine max (soybean)	1,461,722
Triticum aestivum (wheat)	1,286,372
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,480
Oryza sativa (rice)	1,253,557
Ciona intestinalis	1,205,674
Rattus norvegicus + sp. (rat)	1,162,136
Drosophila melanogaster (fruit fly)	821,005
Panicum virgatum (switchgrass)	720,590
Xenopus laevis (African clawed frog)	677,911
Oryzias latipes (Japanese medaka)	666,891
Brassica napus (oilseed rape)	643,881
Gallus gallus (chicken)	600,434
Bombyx mori (domestic silkworm)	568,825
Hordeum vulgare + subsp. vulgare (barley)	501,838
Salmo salar (Atlantic salmon)	498,245
Vitis vinifera (wine grape)	446,664
Caenorhabditis elegans (nematode)	396,687
Phaseolus coccineus	391,150
Porphyridium cruentum	386,903
Canis lupus familiaris (dog)	382,638
Physcomitrella patens subsp. patens	362,131
Ictalurus punctatus (channel catfish)	354,516
Ovis aries (sheep)	338,483
Branchiostoma floridae (Florida lancelet)	334,502
Nicotiana tabacum (tobacco)	334,808
Pinus taeda (loblolly pine)	328,662
Malus x domestica (apple tree)	325,020
Picea glauca (white spruce)	313,110
Aedes aegypti (yellow fever mosquito)	301,596
Gossypium hirsutum (upland cotton)	297,522
Solanum lycopersicum (tomato)	297,142
Oncorhynchus mykiss (rainbow trout)	287,564
Linum usitatissimum	286,852
Macropus eugenii (tammar wallaby)	280,713
Neurospora crassa	277,147
Gasterosteus aculeatus (three spined stickleback)	276,992
Medicago truncatula (barrel medic)	269,501
Pimephales promelas	258,504
Gadus morhua (Atlantic cod)	257,217

dbEST- Datenbank

Achtung:
Neuere RNA-Sequenzierungs-
Projekte (= RNA-Seq)
werden in anderen
Datenbanken wie z.B.
dem SRA abgelegt!

Ein dbEST- Eintrag...

1: AI381340. tc48h12.x1 Soares...[gi:4194121]

IDENTIFIERS

dbEST Id: 2176767
EST name: tc48h12.x1
GenBank Acc: AI381340
GenBank gi: 4194121

Klonnummer;
3 ' read!!

CLONE INFO

Clone Id: IMAGE:2067911 (3')
Source: NCI
Insert length: 819
DNA type: cDNA

PRIMERS

Sequencing: -40UP from Gibco
PolyA Tail: Unknown

SEQUENCE

GCACACAAAACCAAGTTTATTTCTCATGAATTTATAGGACCACAGTCAGCACAAAAGCAAAT
GGTACAAGTGCAAAATGGCTGGGGTGAGGGCGGGGGCCTCTGCTGCTGCCTTCTTTTTTTT
CCAGCTGCCTGAGTACTGGCACACAGCGGGCAGTGAATCTGAGGAGCTGTGGCCTCACAG
TCGCTGCAGGCTGAGCCAAAAAGGAGGGCAGGG

Quality: High quality sequence stops at base: 90

Entry Created: Jan 27 1999
Last Updated: Mar 18 1999

COMMENTS

This clone is available **Gewebe** through LLNL ; contact
the IMAGE Consortium (info@image.llnl.gov) for further
information.

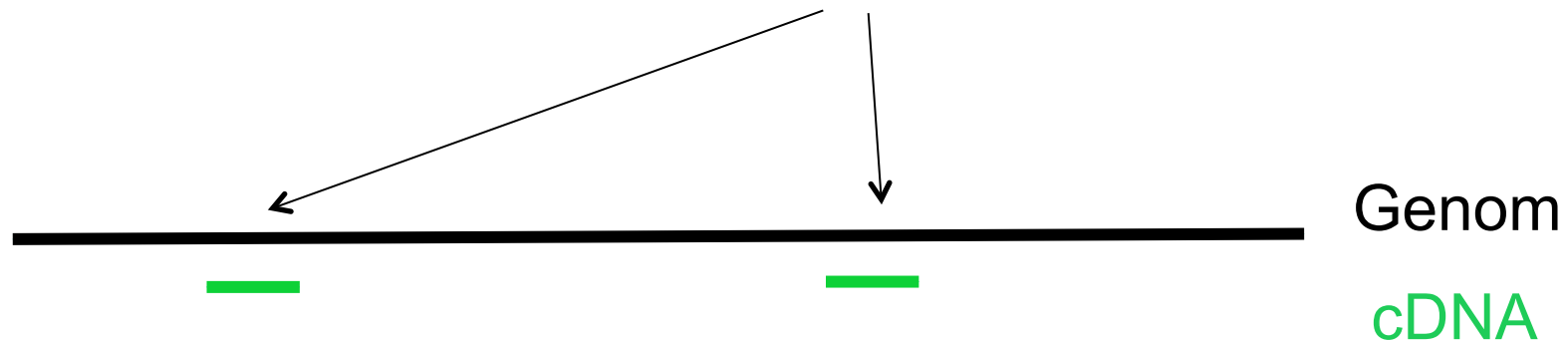
LIBRARY

Lib Name: Soares_total_fetus_Nb2HF8_9w
Organism: [Homo sapiens](#)
Develop. stage: 8-9 weeks
Lab host: DH10B
Vector: pT7T3D-Pac (Pharmacia) with a modified polylinker
R. Site 1: Not I
R. Site 2: Eco RI
Description: 1st strand cDNA was prepared from mRNA obtained from pooled
8-9 week (total) fetus material with a Not I - oligo(dT)
primer [5'
TGTTACCAATCTGAAGTGGGAGCGGCCGCTTAATTTTTTTTTTTTTTTTTT 3'].
Double-stranded cDNA was ligated to Eco RI adaptors
(Pharmacia), digested with Not I and cloned into the Not I
and Eco RI sites of the modified pT7T3 vector. Library went
through one round of normalization, and was constructed by
Bento Soares and M. Fatima Bonaldo.

Herstellung der
cDNA-Bank

Nutzen von cDNA-Datenbanken

- cDNAs, die zu einem Genomabschnitt passen, sind ein erster Beweis, dass dieser Genomabschnitt transkribiert wird!!!!
- cDNA-Sequenzen sind ein phantastisches Hilfsmittel, um per Datenbanksuche **neue Gene zu finden**



EST Profile

Mm.41395 - Ngb: Neuroglobin

Breakdown by Body Sites

Mm.41395		
adipose tissue	0	0 / 1540
adrenal gland	0	0 / 2592
bladder	0	0 / 16283
blood	0	0 / 16776
bone	0	0 / 34066
bone marrow	0	0 / 136333
brain	21	10 / 475384
connective tissue	0	0 / 19807
dorsal root ganglion	82	1 / 12139
embryonic tissue	0	0 / 677554
epididymis	0	0 / 3101
extraembryonic tissue	0	0 / 74703
eye	5	1 / 185387
fertilized ovum	0	0 / 27874
heart	0	0 / 54558
inner ear	0	0 / 37476
intestine	0	0 / 86859
joint	0	0 / 16963
kidney	0	0 / 123578
liver	71	8 / 111370
lung	0	0 / 99799
lymph node	0	0 / 14686
mammary gland	0	0 / 303048
molar	0	0 / 3533
muscle	0	0 / 27159
nasopharynx	0	0 / 7955
olfactory mucosa	0	0 / 3375
ovary	36	2 / 54858
oviduct	0	0 / 3825
pancreas	0	0 / 106229
pineal gland	0	0 / 3906
pituitary gland	0	0 / 18069
prostate	0	0 / 29507
salivary gland	0	0 / 19385
skin	0	0 / 118925
spinal cord	40	1 / 24757
spleen	0	0 / 92417
stomach	0	0 / 31760
sympathetic ganglion	100	1 / 9986
testis	0	0 / 121820
thymus	0	0 / 121153

Nutzen von cDNA-Datenbanken

- die relative Menge von cDNA-Sequenzen eines Gens (zB hier ESTs) geben Auskunft über die ungefähre **Expressionsstärke** des Gens sowie über die Gewebe, in denen das Gen exprimiert wird!*

*Dieser Gedanke wird bei RNA-Seq Daten besonders wichtig, da hierbei ja viel mehr cDNA-Schnipsel sequenziert werden als beim klassischen EST-Ansatz und daher quantitative Aussagen der Genexpression viel genauer sind!

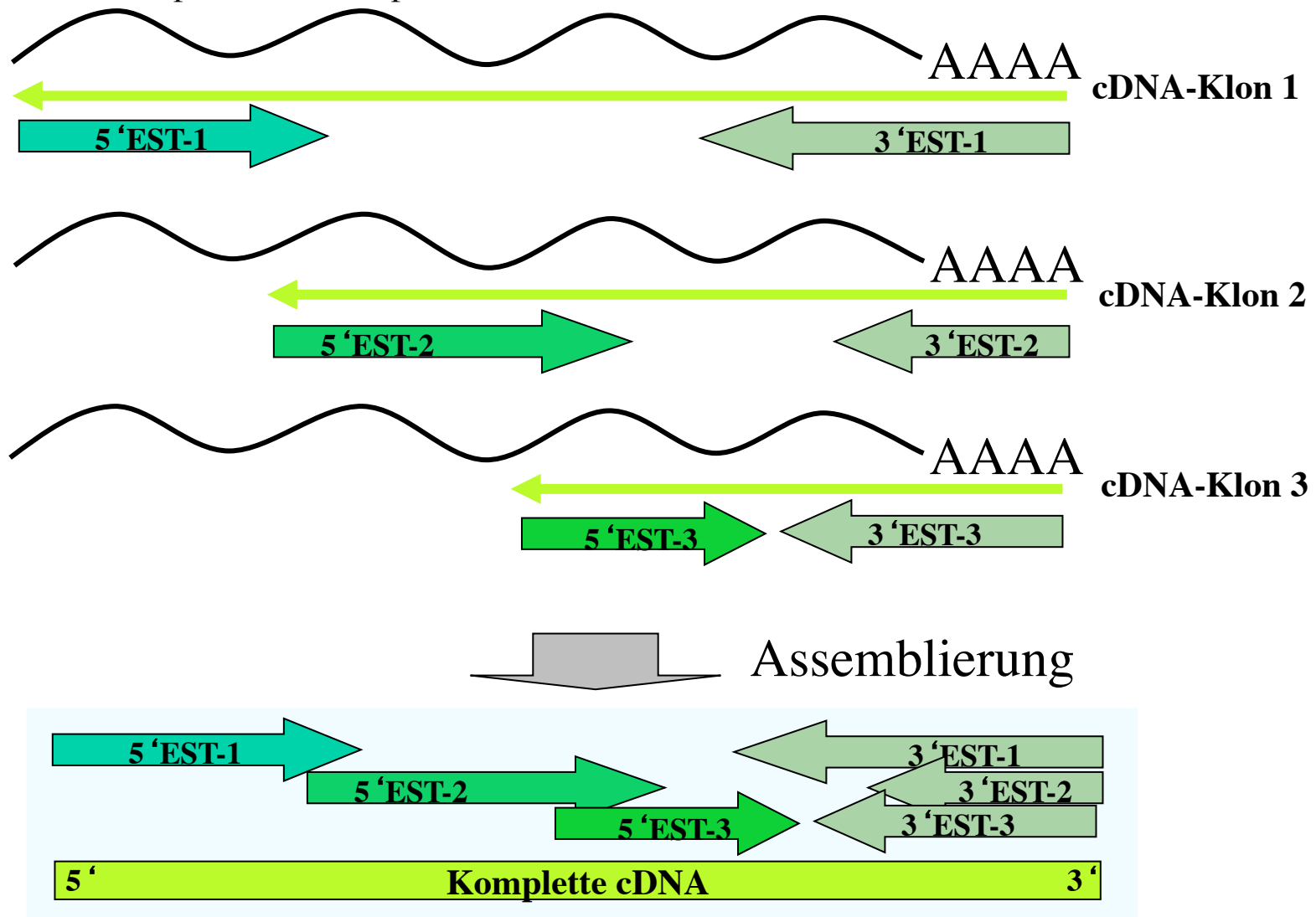
Die dbEST-Datenbank

Aber Achtung!!

- ESTs enthalten viele Lesefehler!
- nicht komplett gespleißte cDNAs
- genomische Verunreinigung
und Artefakte des Klonierungsvorgangs (artifizielle Ligate)

Von ESTs zur kompletten cDNA

Die reverse Transkription ist selten perfekt....



Vom EST zum Gen...

dbEST > Sammlung von cDNA-Schnipseln



„UniGene“-DB > Assemblierung der Schnipsel
führt zu kompletten cDNAs („UniGene-Cluster“)



„Refseq“-DB > Annotator prüft Richtigkeit eines Unigene-Clusters
und erstellt modellhaft oder später verifiziert eine
Referenz-mRNA-Sequenz für das Gen



EntrezGene > alle Infos (mRNA, Protein, Genomsequenz, Position,
Expressionsdaten, Literatur) werden zusammen-
gefasst dargestellt

RefSeq-

„die Referenz unseres Wissensstandes“

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

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

RefSeq-

„die Referenz unseres Wissensstandes“

Key Characteristics of GenBank *versus* RefSeq

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

NCBI Gene - „Alles über mein Gen“

NGB neuroglobin [Homo sapiens (human)] - Gene - NCBI

http://www.ncbi.nlm.nih.gov/gene/58157

UNIMAIL JGU Anmelden Google Phylogeny p... and system ilias 78 molgen UCSC LEO News v ElektrZeitschr

NCBI Resources How To

Gene

Display Settings: Full Report Send to:

NGB neuroglobin [Homo sapiens (human)]

Gene ID: 58157, updated on 2-Nov-2014

Summary

Official Symbol NGB provided by HGNC
Official Full Name neuroglobin provided by HGNC
Primary source HGNC:HGNC:14077
See related Ensembl:ENSG00000165553; HPRD:05602; MIM:605304; Vega:OTTHUMG00000171558
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Summary This gene encodes an oxygen-binding protein that is distantly related to members of the globin gene family. It is highly conserved among other vertebrates. It is expressed in the central and peripheral nervous system where it may be involved in increasing oxygen availability and providing protection under hypoxic/ischemic conditions. [provided by RefSeq, Jul 2008]

Genomic context

Location: 14q24.3 See NGB in Epigenomics, MapViewer
Exon count: 4

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	14	NC_000014.9 (77265491..77271312, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	14	NC_000014.8 (77731834..77737655, complement)

The diagram shows a segment of Chromosome 14 from position 77,158,167 to 77,331,597. The NGB gene is located at approximately 77,265,491 to 77,271,312 (complement). Other features include LOC101926998, TMEM63C, MIR1260A, and PONT2.

GeneCards = Alternative zu EntrezGene

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).

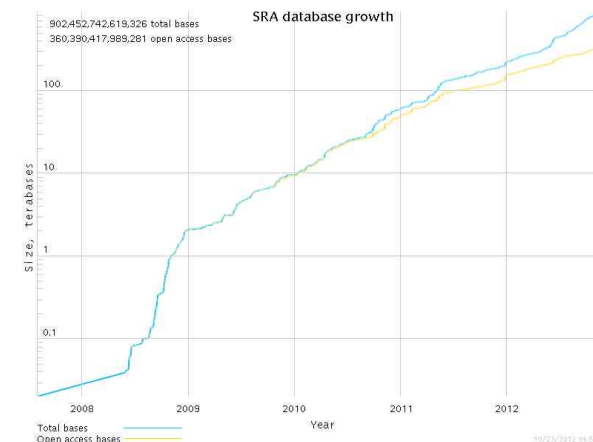
28 clones

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!

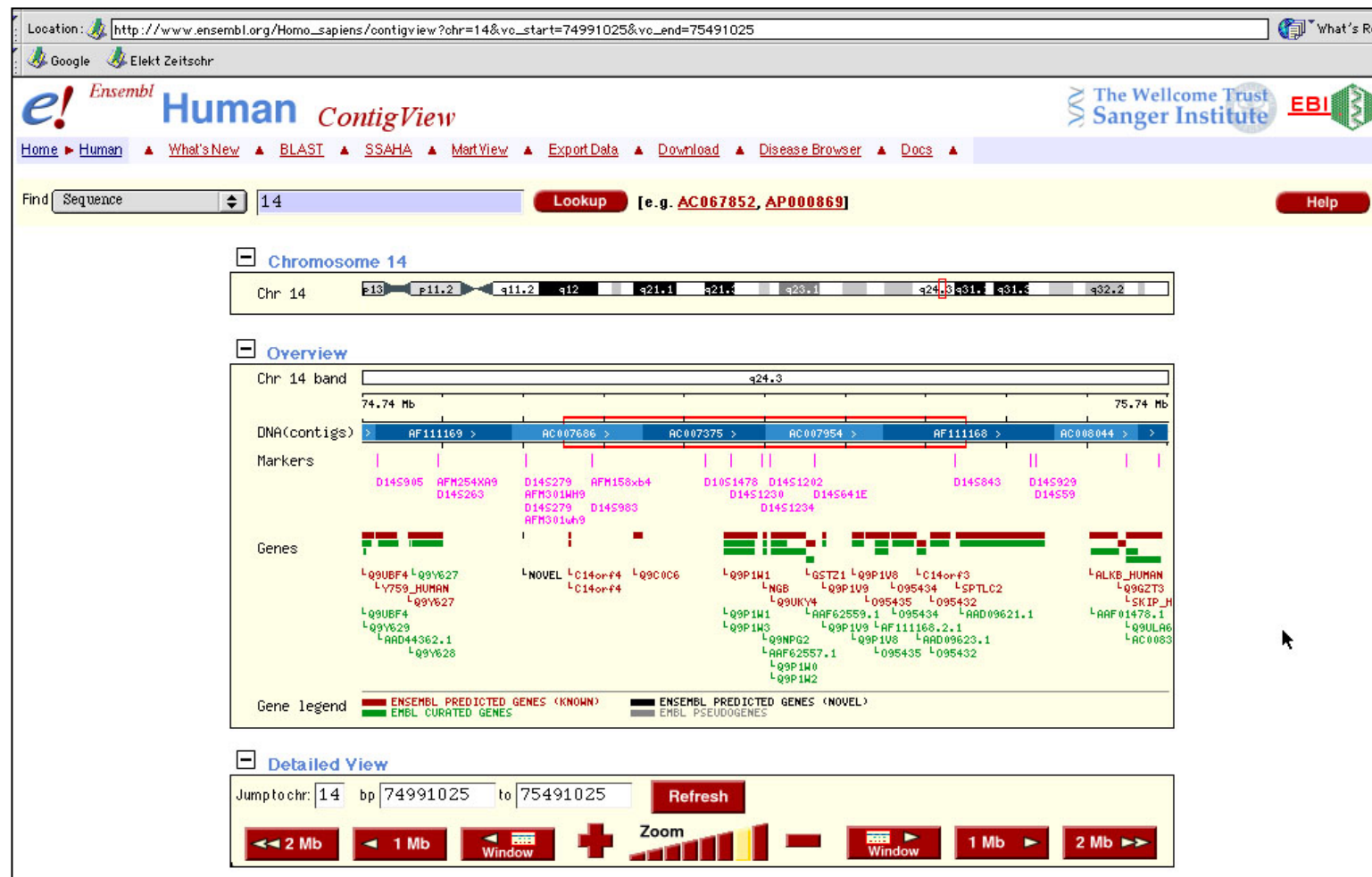
Die Sequenzen hier sind meist unpubliziert...

- **T**race **A**rchive: unannotierte Sanger-Reads aus EST-und Gesamtgenom-Projekten (>500 Spezies, 2.1 Milliarden Reads)
- **S**equen**R**ead **A**rchive: NGS reads diverser Technologien (Tbytes!)



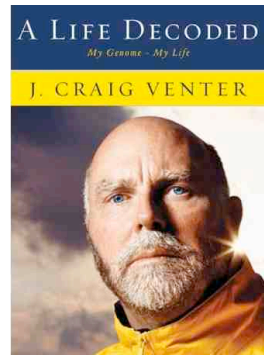
Spezielle annotierte Datenbanken für Genome

www.ensembl.org



Personal Genomes

<http://huref.jcvi.org/>



Examples: [KLKB1^1000](#), [rs2691286](#), [ENSG00000118519](#), [chr19:57550000+40000](#)

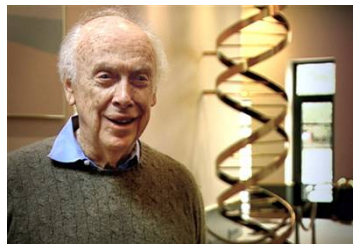
About the HuRef Project

HuRef Genome Browser

This browser enables access to the diploid genome sequence of J. Craig Venter as recently published in [PLoS Biology](#). The graphical interface depicts the haploid sequence with SNP and insertion/deletion DNA variants as identified by genome assembly and comparison methods. The interface also represents the haplotype blocks from which diploid genome sequence can be inferred and gene annotations. This work was done at the J. Craig Venter Institute, with collaborators from The Hospital for Sick Children in Toronto, Canada, University of California, San Diego, and Universidad de Barcelona, Spain.

Publications

- [The HuRef Browser: a web resource for individual human genomics](#) Axelrod N, Lin Y, Ng PC, Stockwell TB, et al.
- [The Diploid Genome Sequence of an Individual Human](#) Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al.
- [The Diploid Genome Sequence of an Individual Human](#) (supplemental poster).
- [A New Human Genome Sequence Paves the Way for Individualized Genomics](#) by Liza Gross.



James Watson's Personal Genome Sequence

README: How do I use the James Watson Genome Browser?
Downloads: Download bulk JW polymorphisms. For the complete data set, please go to the [NCBI Trace Archive](#) and search for *CENTER_NAME* = 'CSHL' and *CENTER_PROJECT* = 'Project Jim'.

Darstellung 34.46 kbp von chr7, Position 75,221,807 bis 75,256,264

☐ **Anleitungen**
 Suche nach Sequenz Namen, Gen Namen, Locus Namen oder anderen Landmarks. Der Platzhalter * ist erlaubt. Um einen Locus zu zentrieren, auf das Lineal klicken. Um zu vergrößern oder die Position zu verändern verwendet man die Scroll/Zoom Knöpfe.

Beispiele: [HTR2A](#), [macular degeneration](#), [rs726455](#), [DAOA](#), [chr22:20230140..20330139](#), [PARK3](#), [SNP:rs131693](#), [SPTB](#), [NM_001008496](#), [3q21.2](#), [ENM010](#).

[Banner ausblenden] [Bookmark für diese Ansicht] [Link zur Abbildung dieser Ansicht] [Abbildung in Publikationsqualität] [Hilfe] [Reset]

☐ **Suche**

Landmark oder Region:

Daten Quelle

☐ **Überblick**

Dumps, Suchen und andere Operationen:

Scroll/Zoom:

Anzeigen 34.46 kbp

☐ **Ideogram**

<http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/>

Weitere sehr wichtige abgeleitete Datenbanken...

Protein-Familien, -Domänen, 3D-Struktur:

PFAM pfam.wustl.edu/

PROSITE www.expasy.org/prosite/

PDB www.rcsb.org/pdb/

} Diese und andere zusammengefasst in
InterPro (via EBI)

Stoffwechselwege, Enzyme:

KEGG www.genome.ad.jp/kegg/

ENZYME <http://www.expasy.org/enzyme/>

Biochemical Pathways www.expasy.org/cgi-bin/search-biochem-index

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

[Ftp://iubio.bio.indiana.edu/soft/molbio/readseq/](ftp://iubio.bio.indiana.edu/soft/molbio/readseq/)

Seqverter

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

Für Programmierer: Bibliotheken für Fileformate etc sind in
Modul-Sammlungen wie z.B. BioPerl vorhanden