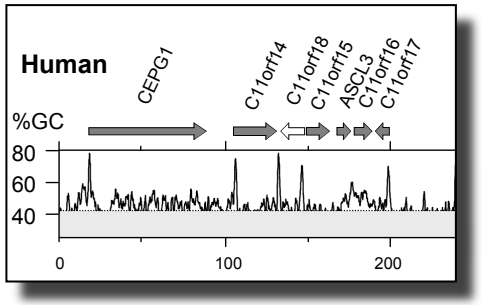
WS 2018/2019

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

Thomas Hankeln



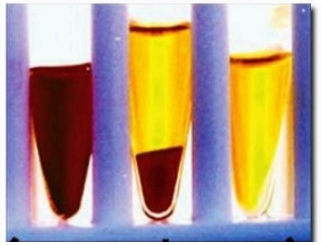
Strategien der Gensuche

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

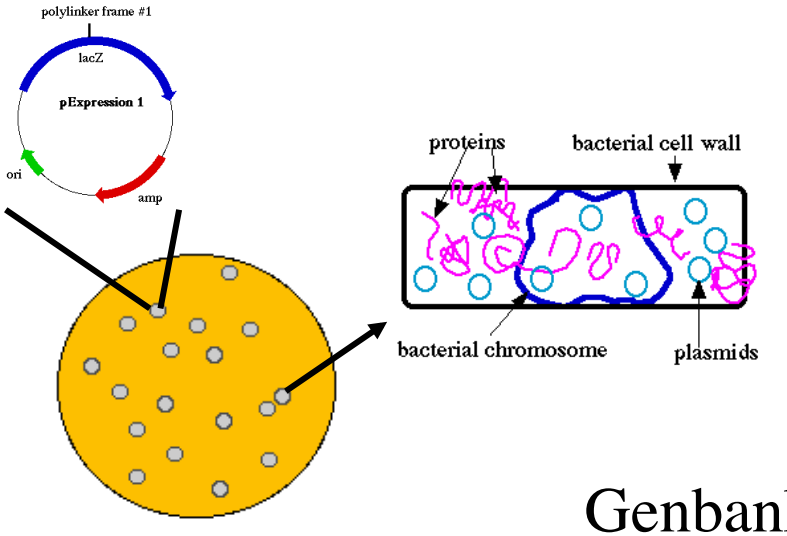
↓

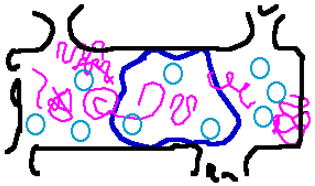
„Screenen“ einer cDNA-Expressions-klonbank mit **AK als Sonde**

↓

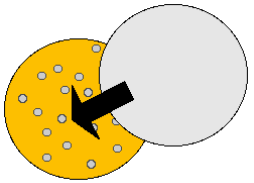
Identifizierung des Klons, der die cDNA des gesuchten Gens enthält und daher das gesuchte Protein bildet

Der biochemische Weg





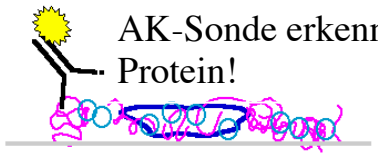
Lyse



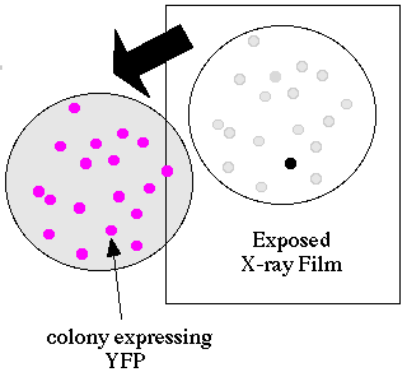
„Abklatsch“
auf Membran

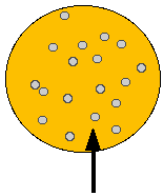
Genbank:

jeder Bakterienklon enthält und exprimiert ein anderes Gen



AK-Sonde erkennt Protein!





Der Klon enthält das gesuchte Gen!!

Der genetische Weg:

Positionelle Klonierung

Voraussetzung:

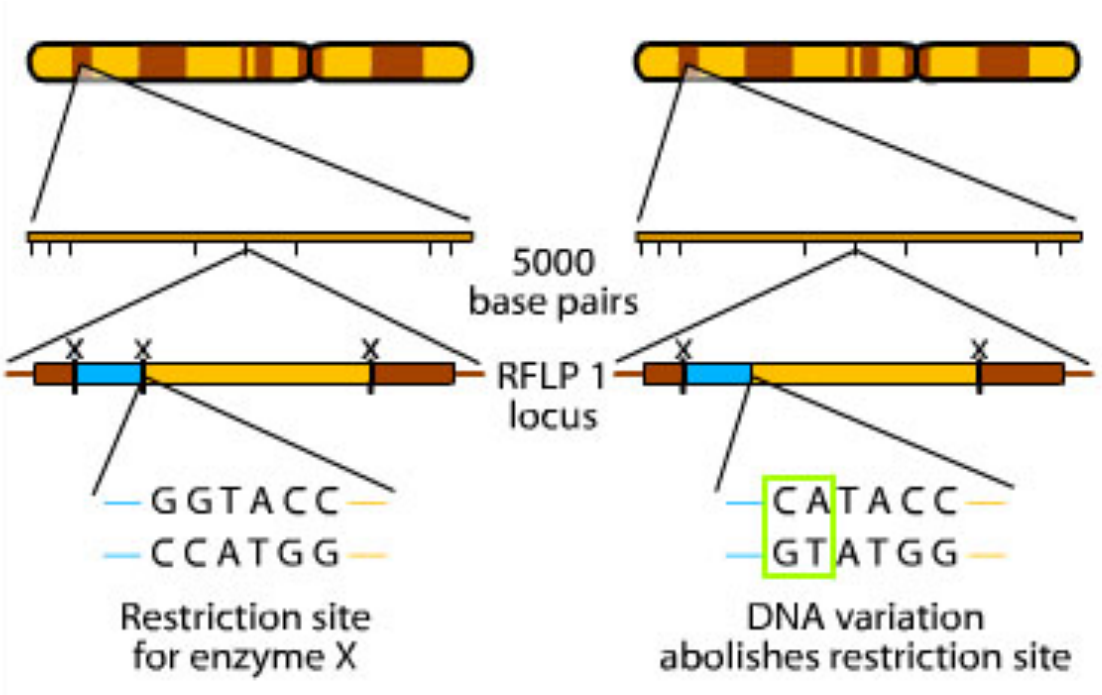
Information über die Lage eines (Krankheits)gens auf den Chromosomen bzw. im Genom

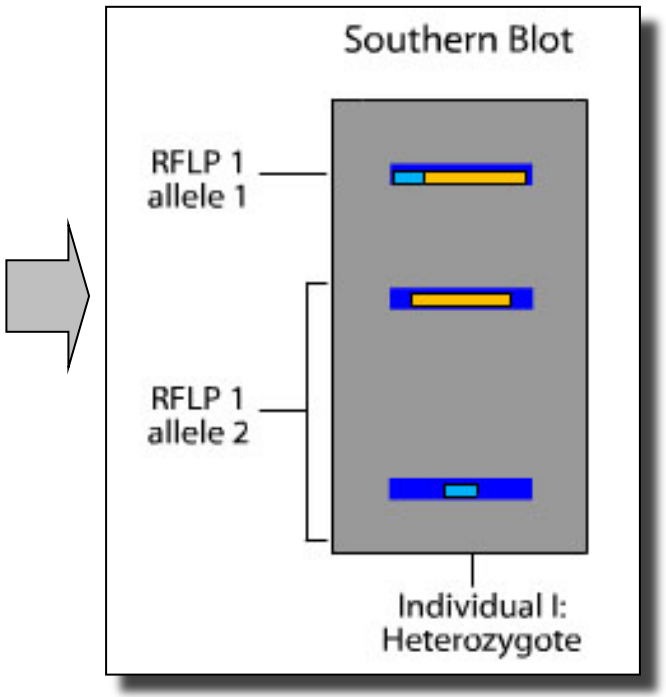
- sichtbare Chromosomen-Anomalien?
- Verlust der Heterozygotie (LOH) in einer Genomregion? (>Tumor-Suppressorgene)
- **Kopplungsanalyse:**

  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





RFLP

(Restriktionsfragment-Längenpolymorphismus)

Positionelle Klonierung:

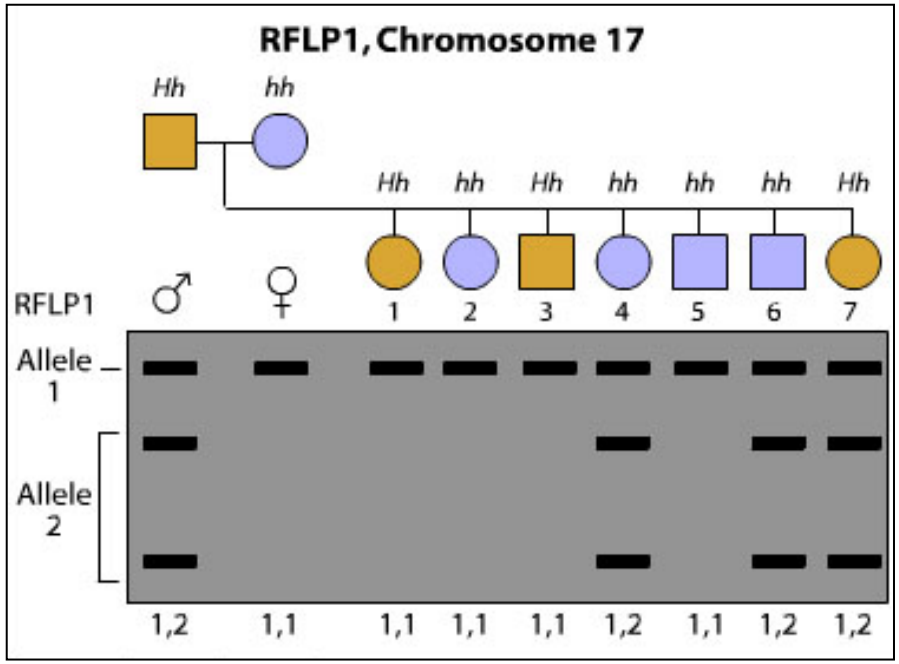
weitere gebräuchliche Markertypen

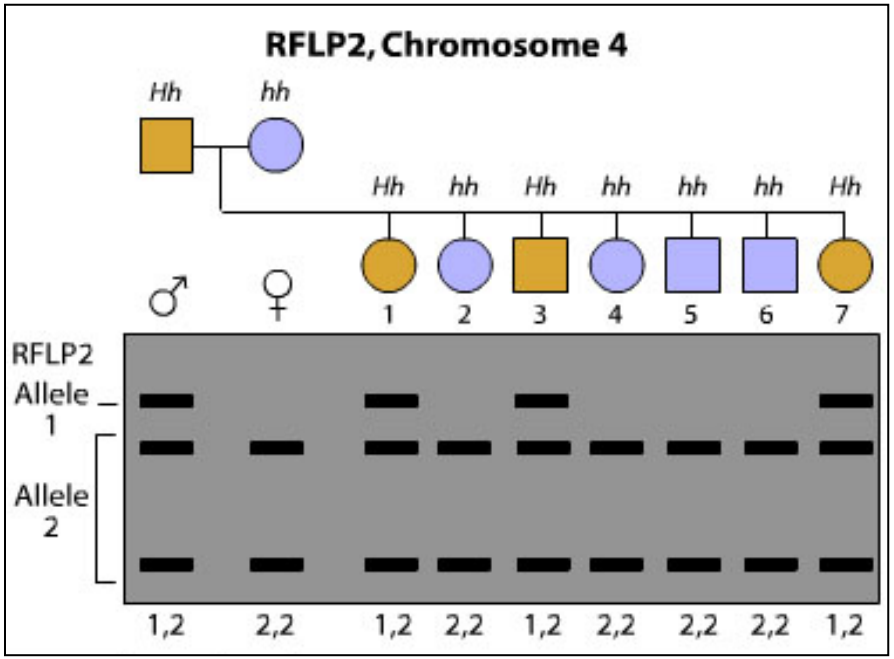
- **Mikrosatelliten** > Detektion per PCR & Gelelektrophorese

- **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!

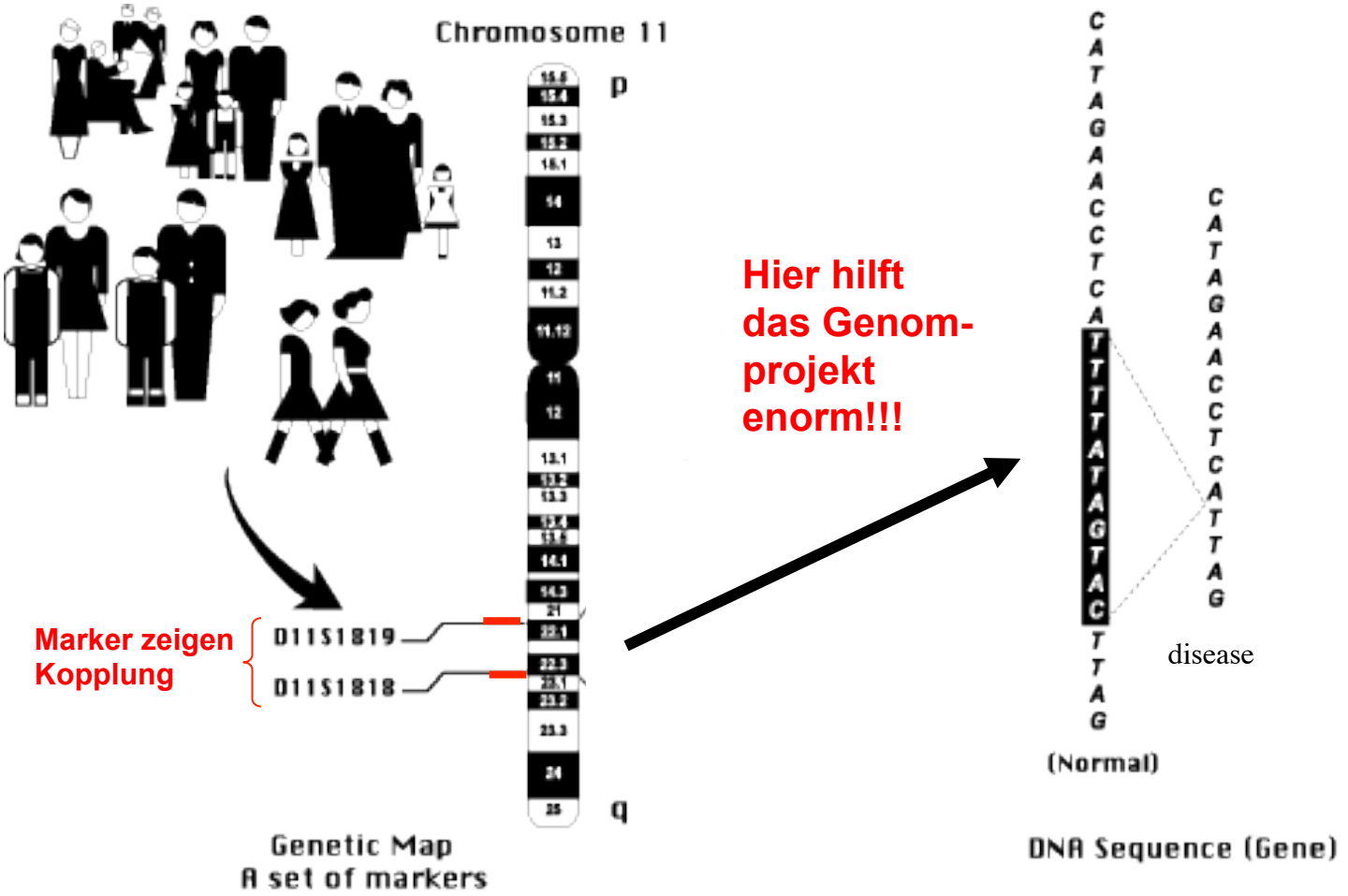
Positionelle Klonierung und Genomsequenzierung ergänzen sich!

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

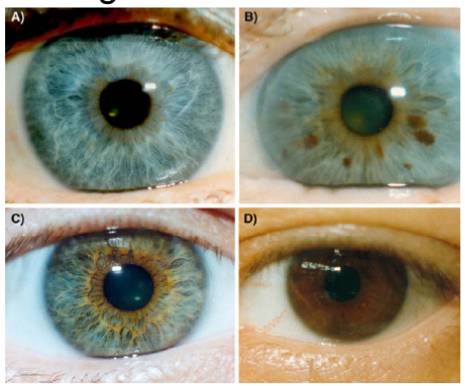


Table 4 Conservation of the eye color silencer sequence in the *HERC2* intron 86 in 9 different species

Species	Eye color	DNA-Library	DNA sequence
Homo	Blue	hg18_dna	TTCATTTGAGCATTAAGTGTCAAGTTCTGCACGCTAT
Homo	Brown	hg18_dna	TTCATTTGAGCATTAAATGTCAAGTTCTGCACGCTAT
Chimpanse	Brown	panTro2_dna	TTCATTTGAGCATTAAATGTCAAGTTCTGCACGCTAT
Rhesus monkey	Brown	rheMac2_dna	TTCATTTGAGCATTAAATGTCAAGTTCTGCACGCTAT
Horse	Brown	equCab1_dna	TTCACTTGCACGCTAAATGTCAAGTGCTGCACAATGT
Cow	Brown	bosTau2_dna	TTCATCTGCATGGTAAATGTCAAGTAC-ACACACTGT
Cat	Brown-yellow	felCat3_dna	TTCATTTGCATGTTAAATGTCAAGTACCACACAATAC
Dog	Brown-yellow	canFam2_dna	TTCATTTGCATGTTAAATGTCAAGTGC-ACACAATAT
Rat	Brown	rn4_dna	TTCATTTGCGTATTAAATGTCAA
Mouse	Brown	Mm8_dna	TTCATTTGCGTATTAAATGTCAAATGCCATGCACTAT
Consensus sequence - blue eye			Ttca-ttg-----taaGtgtcaa-t-c----c-tat
Consensus sequence - brown eye			Ttca-ttg-----taaAtgtcaa-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

Kopplunganalyse 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)

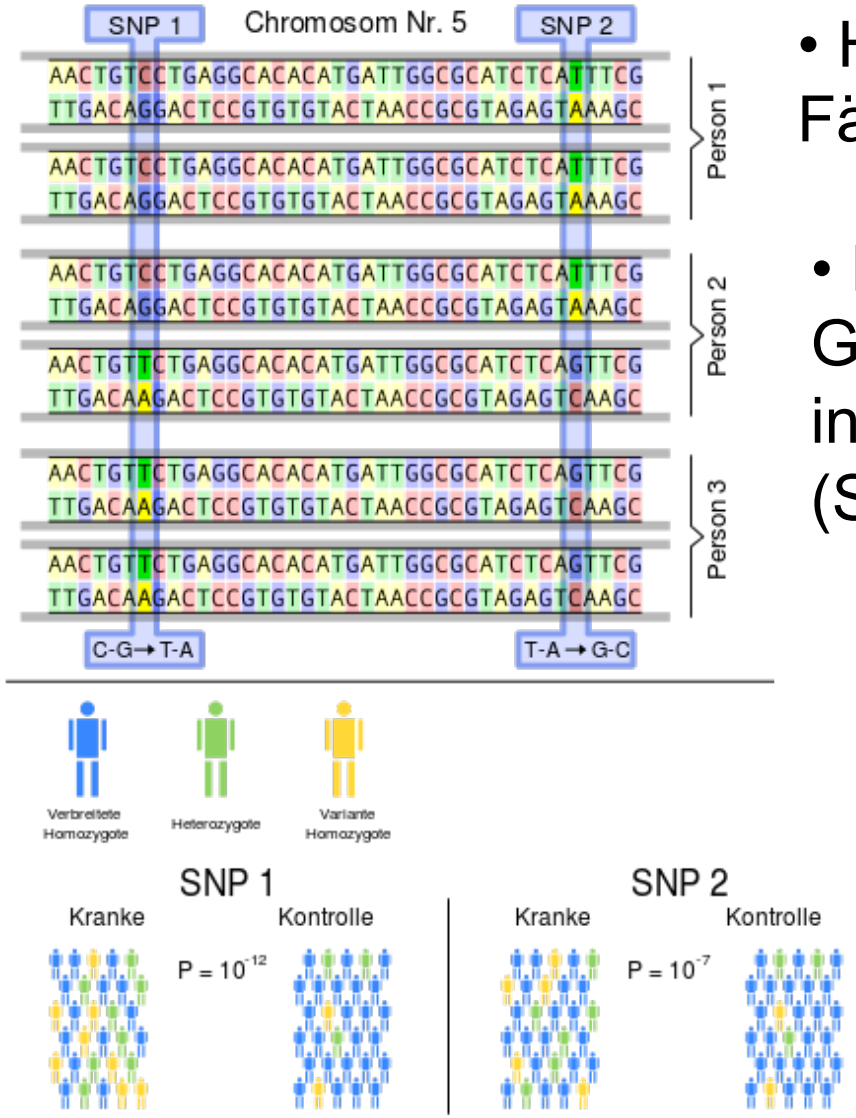
- sind bestimmte genetische Unterschiede (SNPs)
in der Kohorte der Fälle (mit Erkrankung) gegenüber den Kontrollen (Gesunde) gehäuft vorhanden und somit
mit der Erkrankung statistisch signifikant assoziiert?

- 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.

https://de.wikipedia.org/wiki/Genomweite_Assoziationsstudie

Genomweite Assoziationsstudien



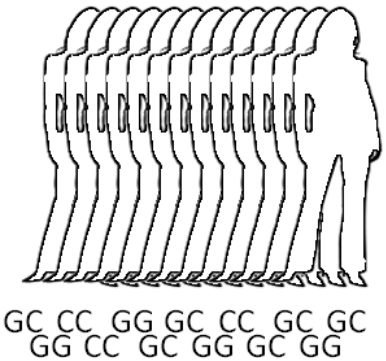
- Hunderte bis Tausende Fälle und Kontrollpersonen > DNA

- 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



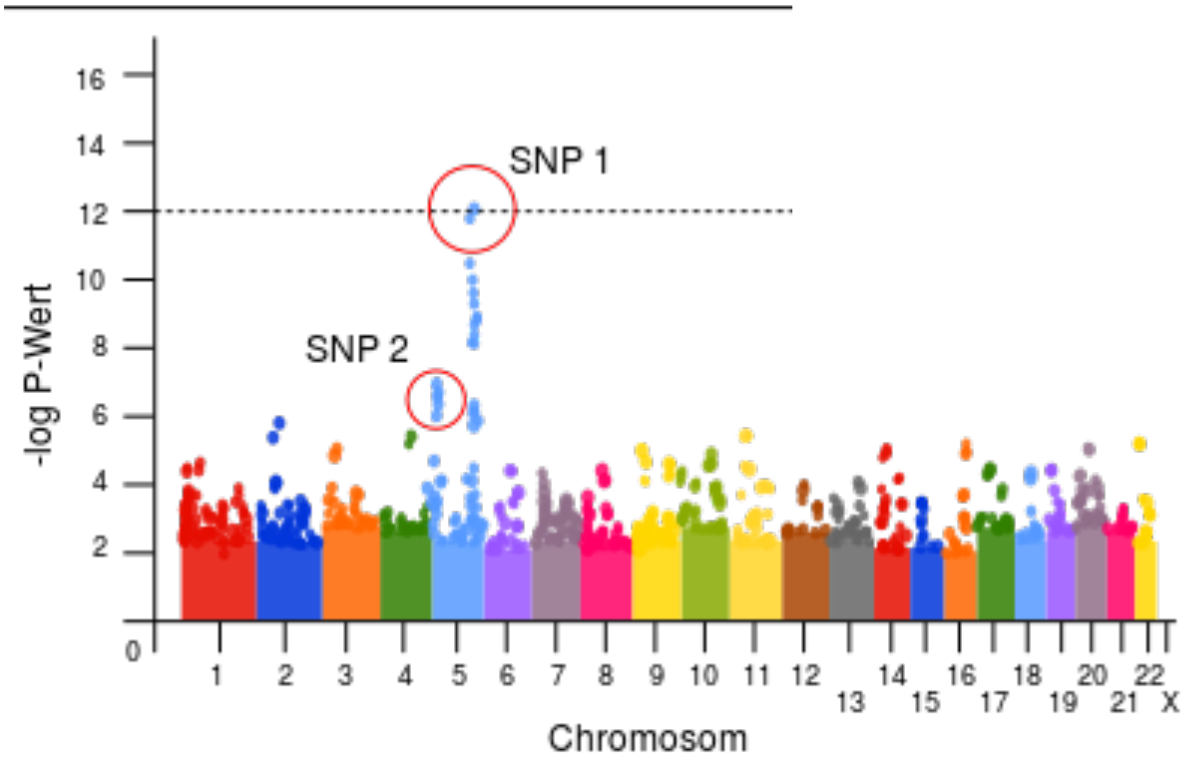
SNP1	*SNP2*	*SNP…*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

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

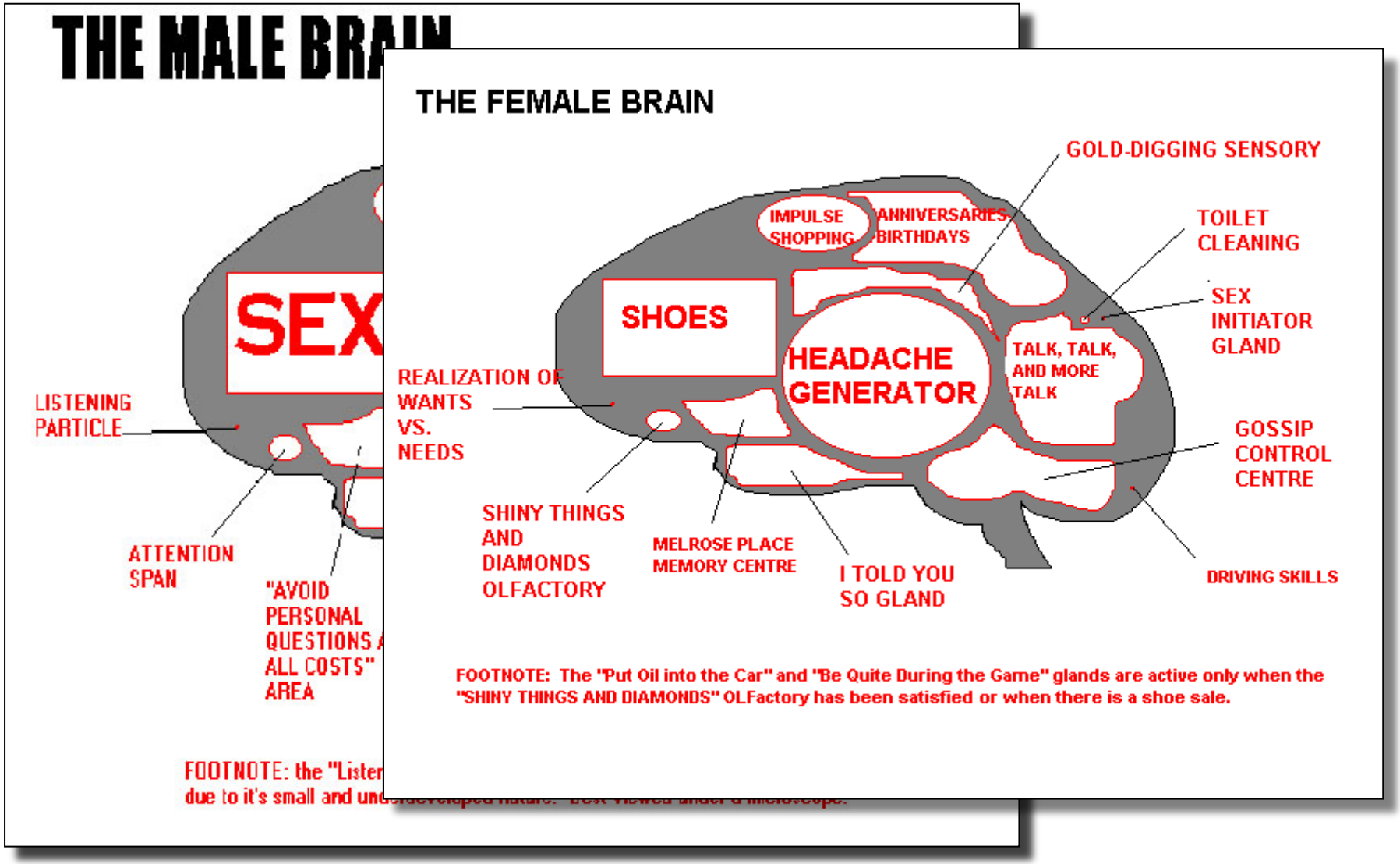
Wege zu „allen Genen“ durch Genomanalyse (Genomik)

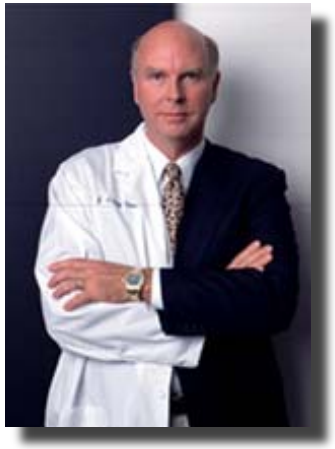
- **Gesamtgenom-Sequenzierung**
  „whole genome shotgun“

- **EST-Sequenzierung**
  > „expressed sequence tags“, cDNA-‘Schnipsel‘
  > Gen-Entschlüsselung ‚leicht gemacht‘
  > genannt **RNA-Seq** bei Verwendung von NGS-Technik

Diese beiden Ansätze komplementieren einander…

Welche Gene sind da aktiv…?

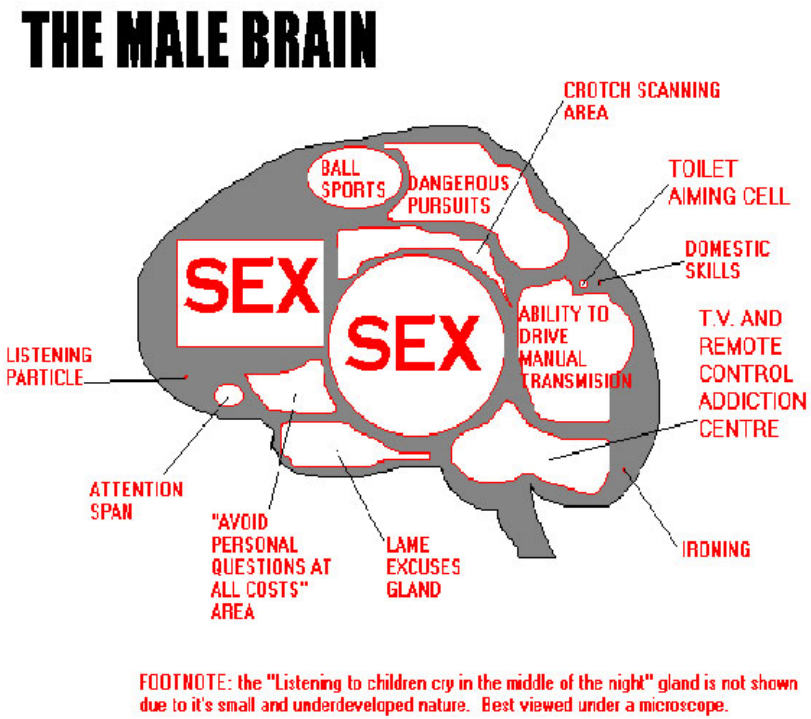




J. C. Venter

EST-Sequenzierung

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



Gewebe (z.B. Gehirn)

↓

RNA-Präparation

↓

Umschreiben in cDNA

↓

(cDNA-Klonbank)

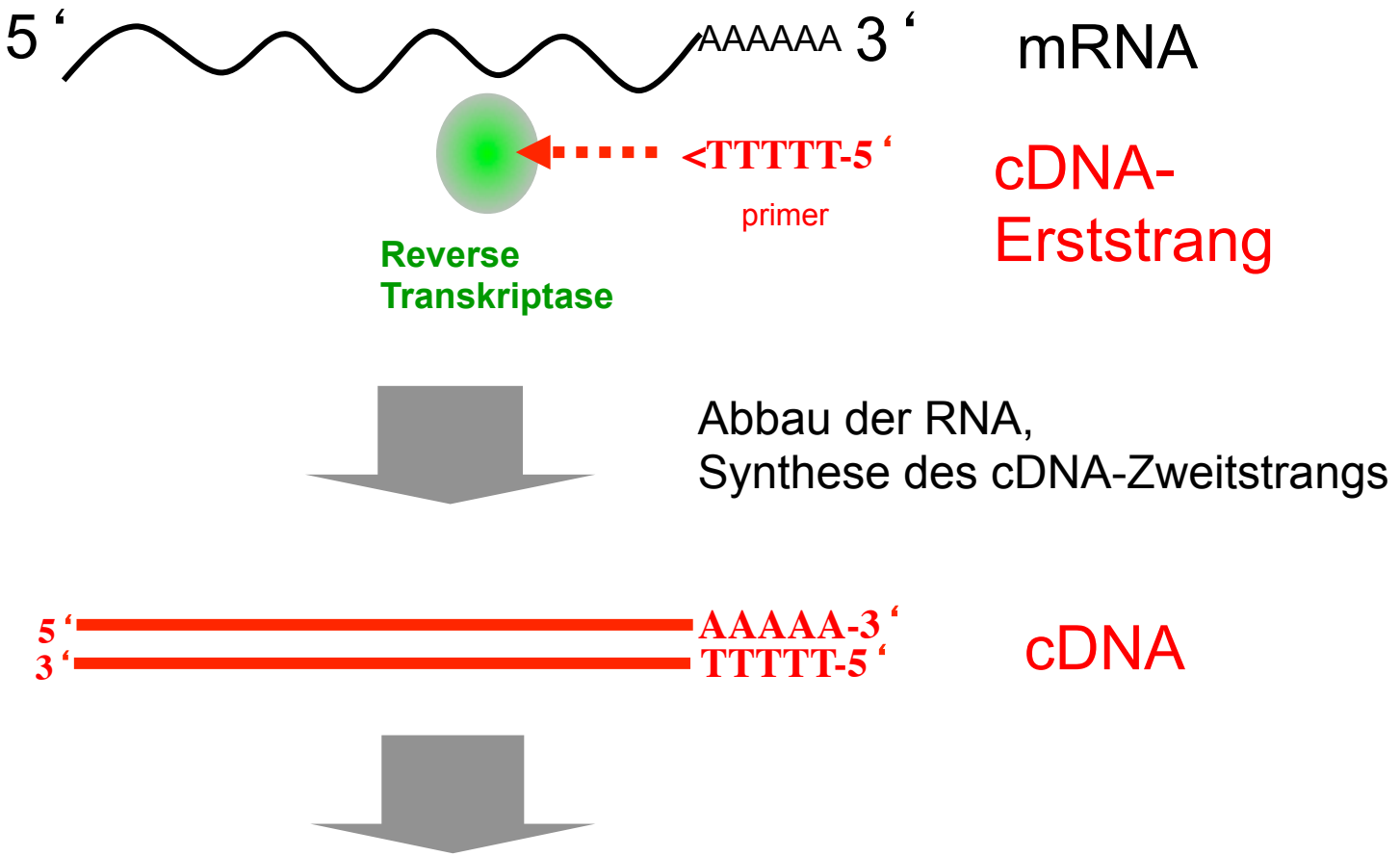
↓

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

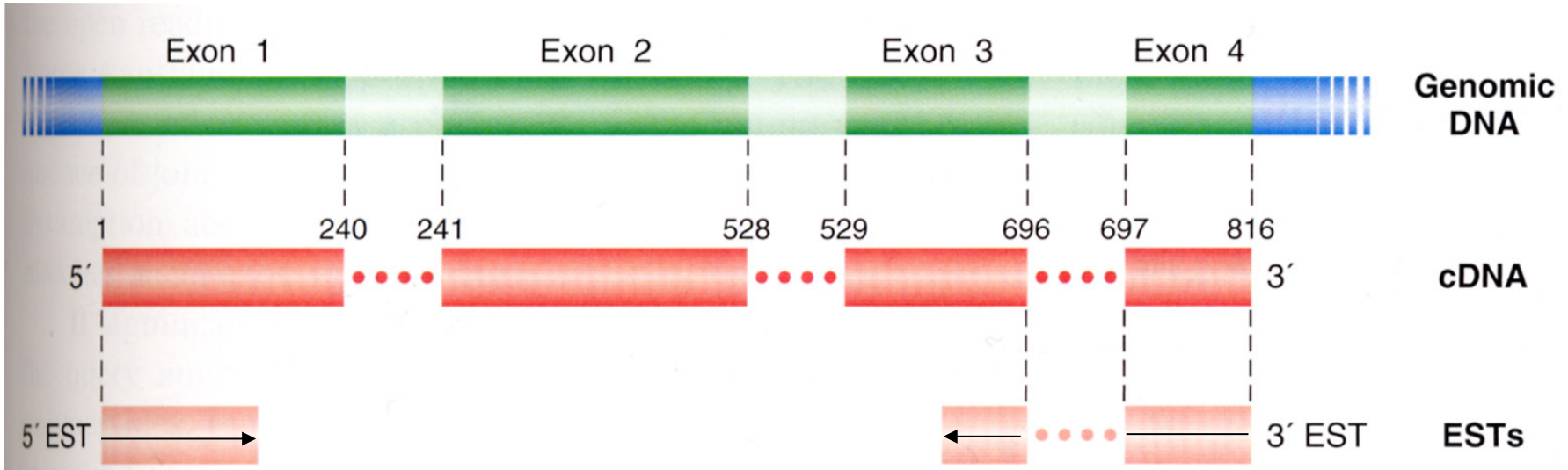


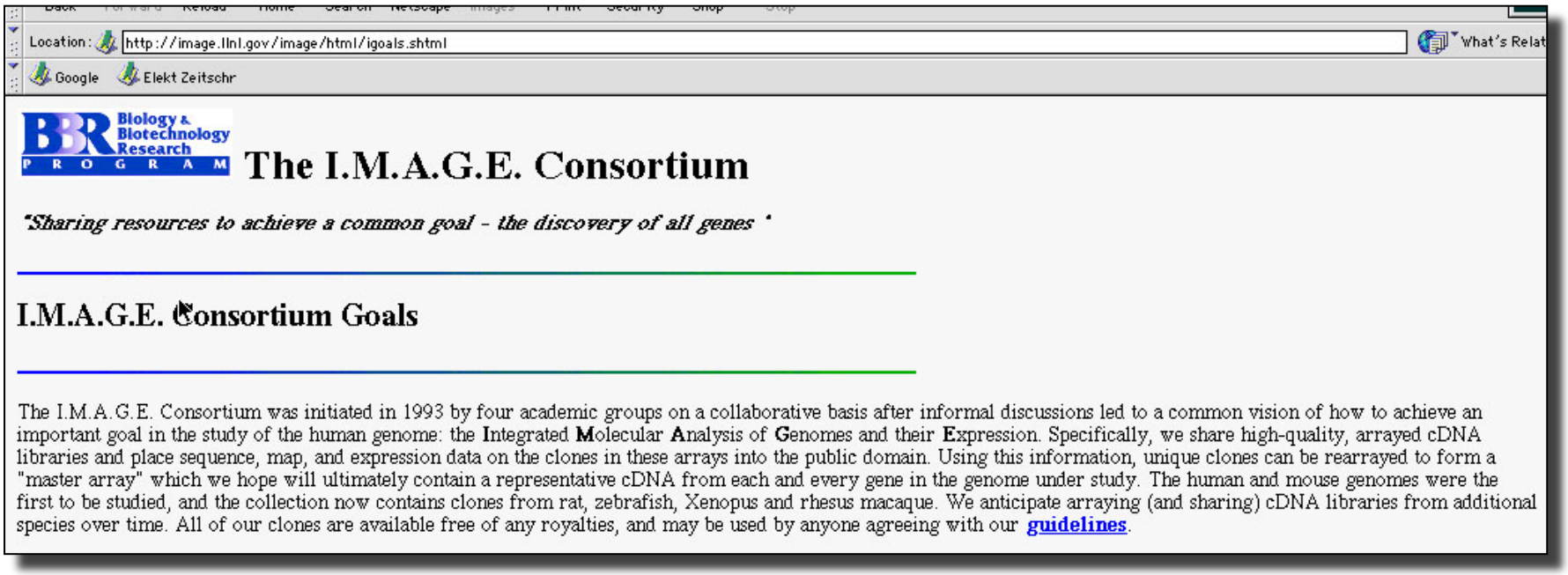
Figure 9-32 The alignment of fully sequenced cDNAs and ESTs with genomic DNA. The solid lines indicate regions of alignment; for the cDNA, these are the exons of the gene. The dots between segments of cDNA or ESTs indicate regions in the genomic DNA that do not align with cDNA or EST sequences; these 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?

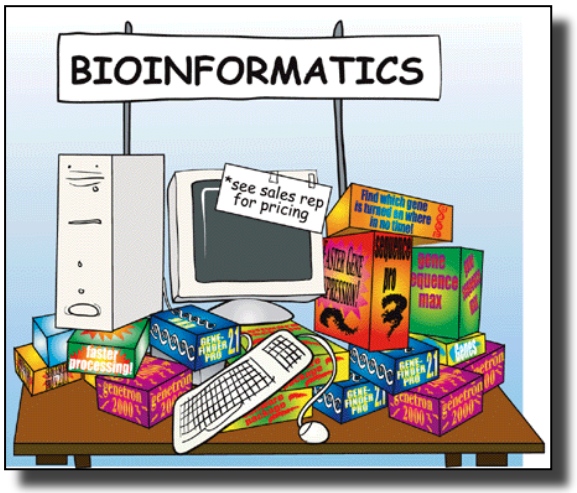
EST-Projekte



- 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

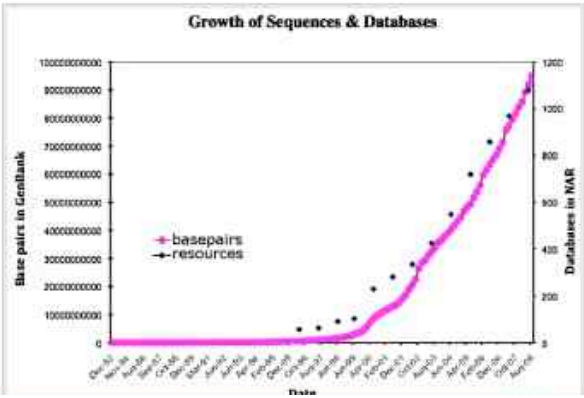
Genomic Data Resources: Challenges and Promises

By: Warren C. Lathe III (*OpenHelix*), Jennifer M. Williams (*OpenHelix*), Mary E. Mangan (*OpenHelix*) & Donna Karolchik (*University of California, Santa Cruz Genome Bioinformatics Group*)
Citation: Lathe, W., Williams, J., Mangan, M. & Karolchik, D. (2008) Genomic Data Resources: Challenges and Promises. *Nature Education* 1(3)

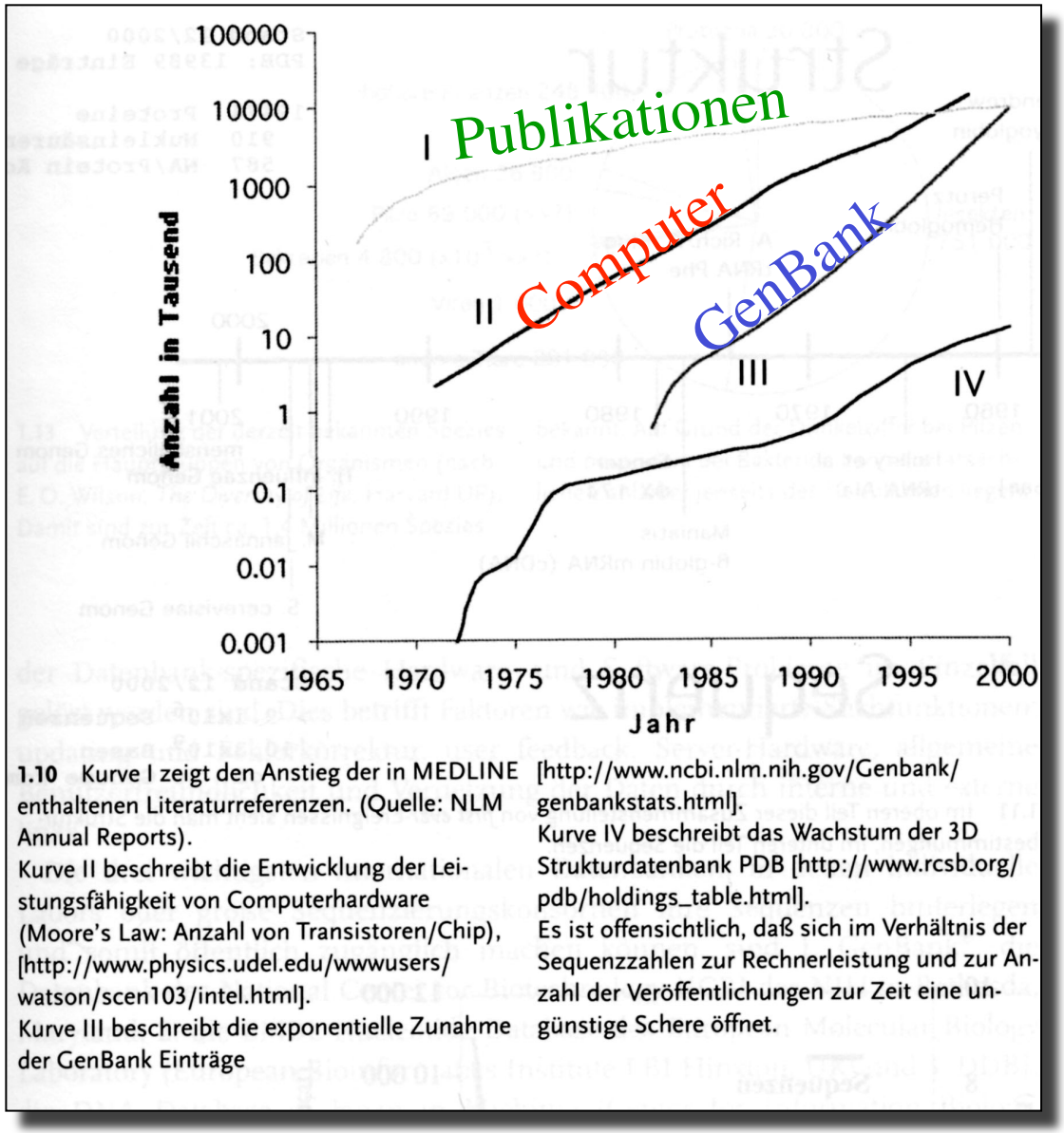
Where would you go to find the nucleotide and amino acid data you need? There are thousands of genomic databases, tools, and other resources freely accessible on the Internet.

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

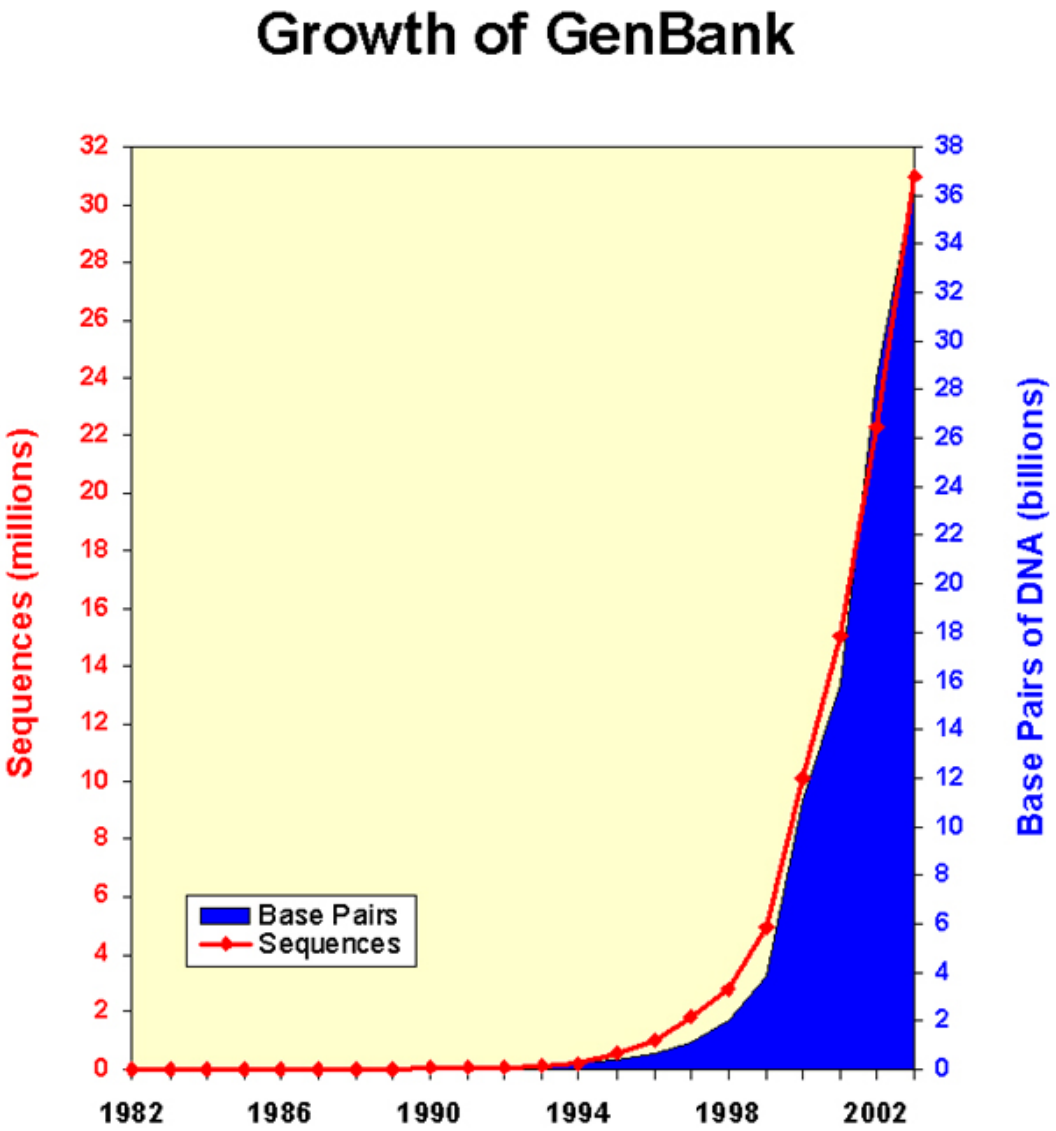


Datenbanken-Wachstum



1.10 Kurve I zeigt den Anstieg der in MEDLINE enthaltenen Literaturreferenzen. (Quelle: NLM Annual Reports).
Kurve II beschreibt die Entwicklung der Leistungsfähigkeit von Computerhardware (Moore's Law: Anzahl von Transistoren/Chip), [http://www.physics.udel.edu/wwwusers/watson/scen103/intel.html].
Kurve III beschreibt die exponentielle Zunahme der GenBank Einträge [http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html].
Kurve IV beschreibt das Wachstum der 3D Strukturdatenbank PDB [http://www.rcsb.org/pdb/holdings_table.html].
Es ist offensichtlich, daß sich im Verhältnis der Sequenzzahlen zur Rechnerleistung und zur Anzahl der Veröffentlichungen zur Zeit eine ungünstige Schere öffnet.

Datenbank-Wachstum



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

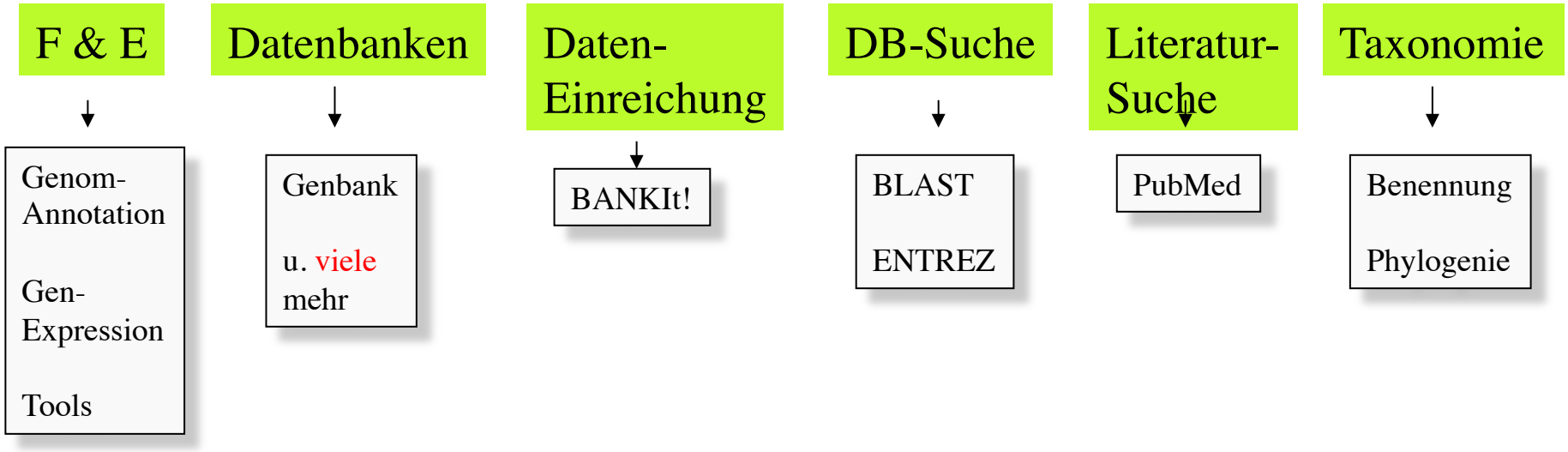
http://www.ncbi.nlm.nih.gov/

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

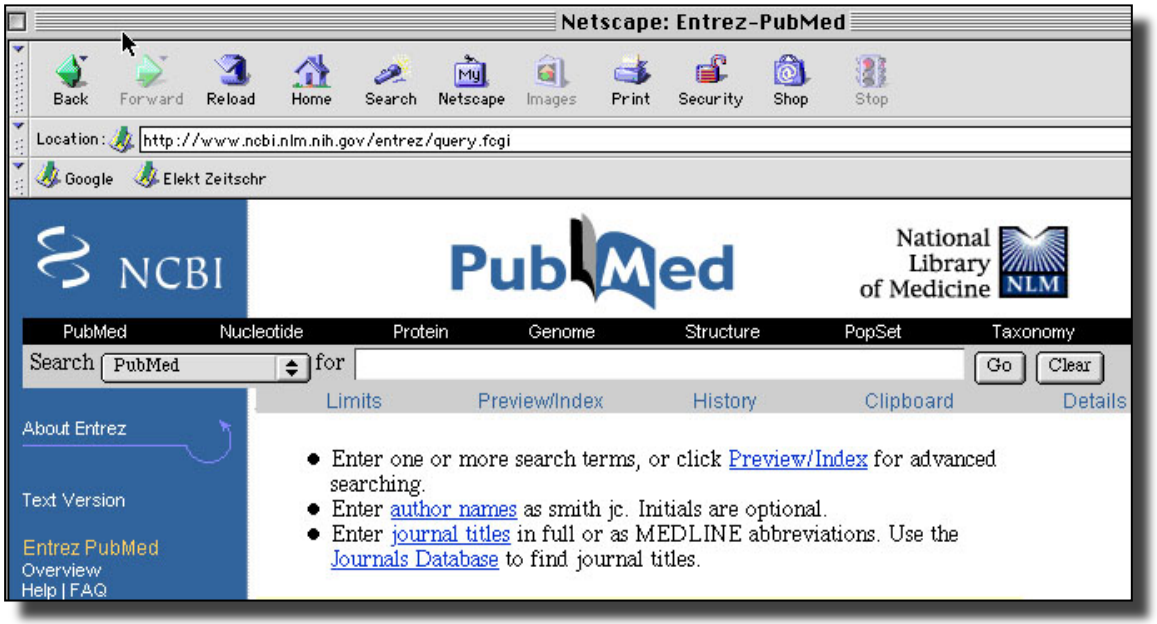


...sowie Trainings, Tutorials und vieles mehr“!!

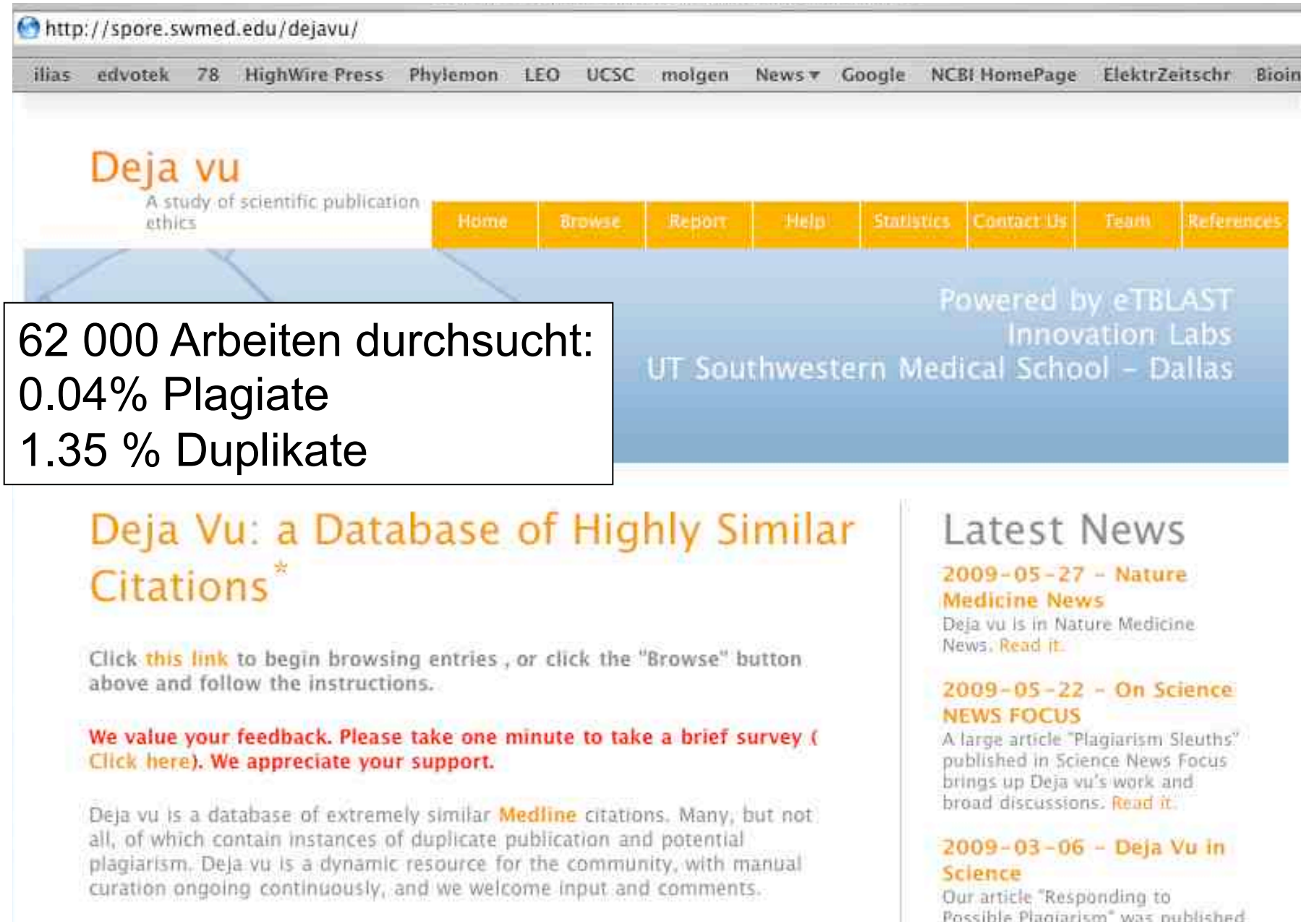
https://www.ncbi.nlm.nih.gov/pubmed/

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: „*“


http://spore.swmed.edu/dejavu/
ilias edvotek 78 HighWire Press Phylemon LEO UCSC molgen News Google NCBI HomePage ElektrZeitschr Bioin
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

https://www.omim.org/about

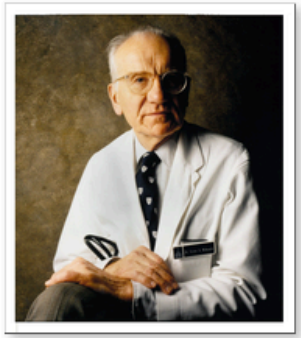
Suchen

eistbesucht news LEO Wikipedia Google UniMail NCBI NTV süd-D ZEIT UCSC molgen Aktuelle Nachrichte... Uni Mainz fileahre Inbox - Outlook

About Statistics Downloads Contact Us MIMmatch Donate Help

Search OMIM... Options

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

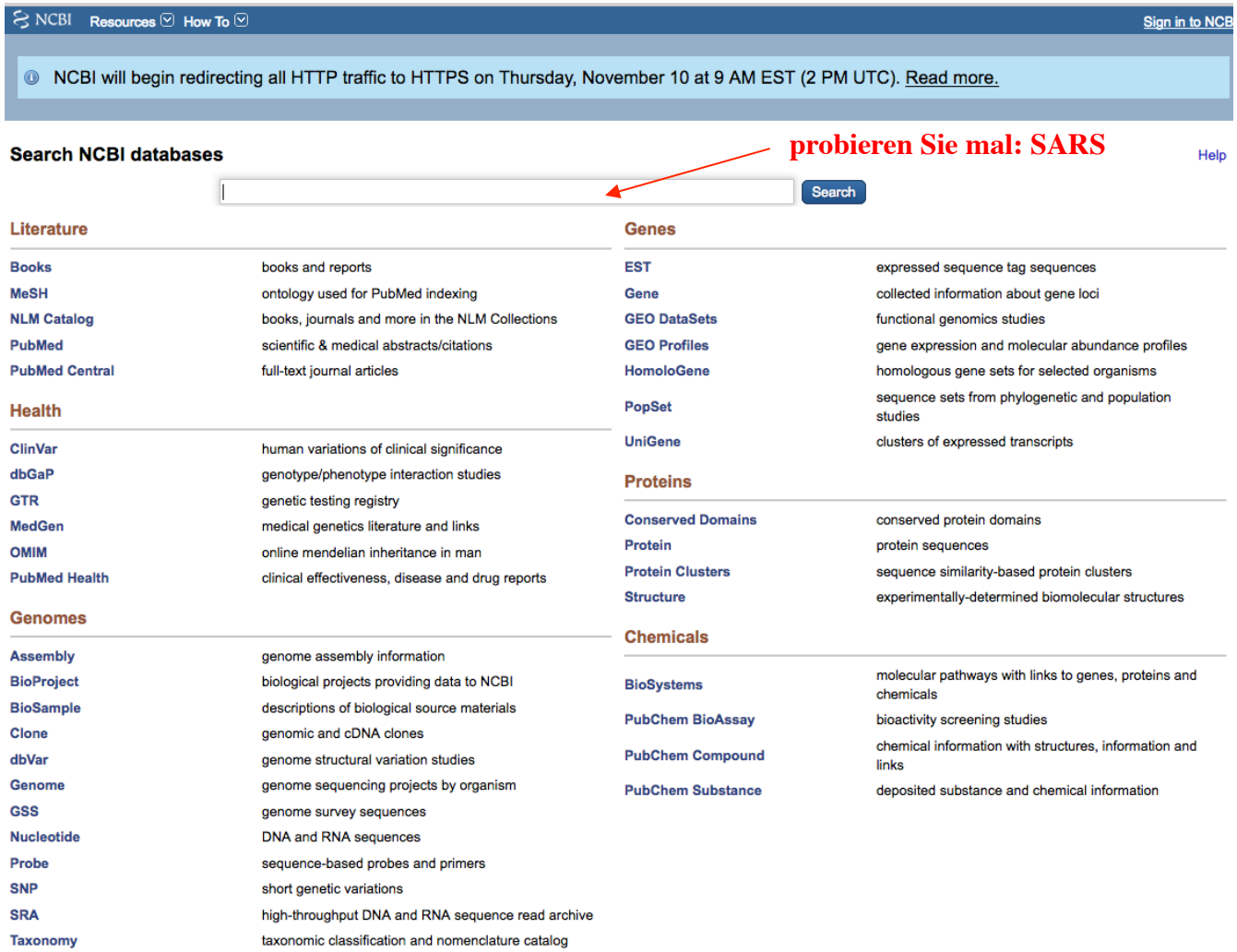
Online Mendelian Inheritance of Man
= Katalog menschlicher Gene und ihrer Erkrankungen

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!



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

…und wie komme ich zu meiner Sequenz?

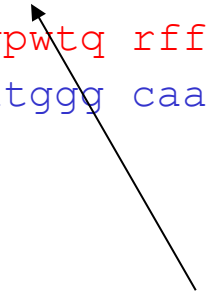
Ich kenne eine Accession Number NM_000518

Ich kenne ein Gensymbol HBB

(Hämoglobin Beta)

Ich kenne einen passenden „Sequenz-Schnipsel“

mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq 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“

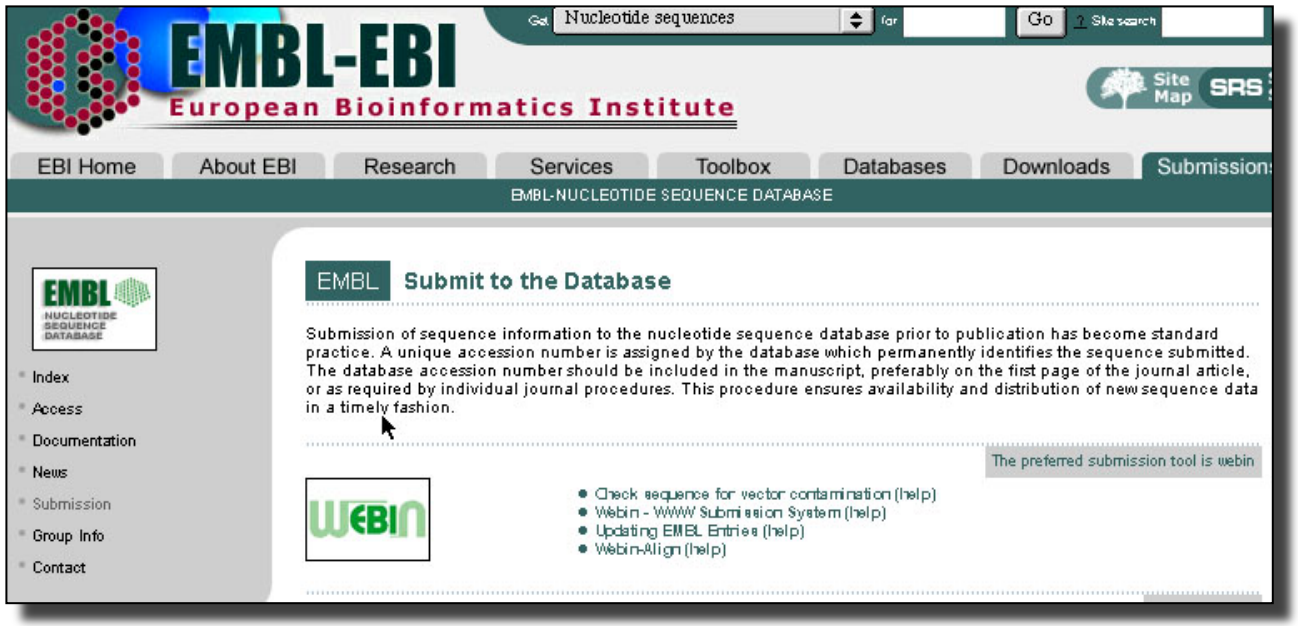
• www.oxfordjournals.org/nar/database/a/
• > 1500 Datenbanken in der Sammlung!!

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

- 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!

Protein-Sequenzdatenbanken

UniProt (via EBI)

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

- ehemals größte Proteindatenbank (250 000 Einträge)
- annotiert, z. T. 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!
- ca. 560 000 Einträge

trEMBL („translated EMBL“)

- automatisch annotiert, nicht curated; > 126 Mio. Einträge

PDB („Protein Data Bank“)

- bekannte 3D-Strukturen

NR (via NCBI)

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

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

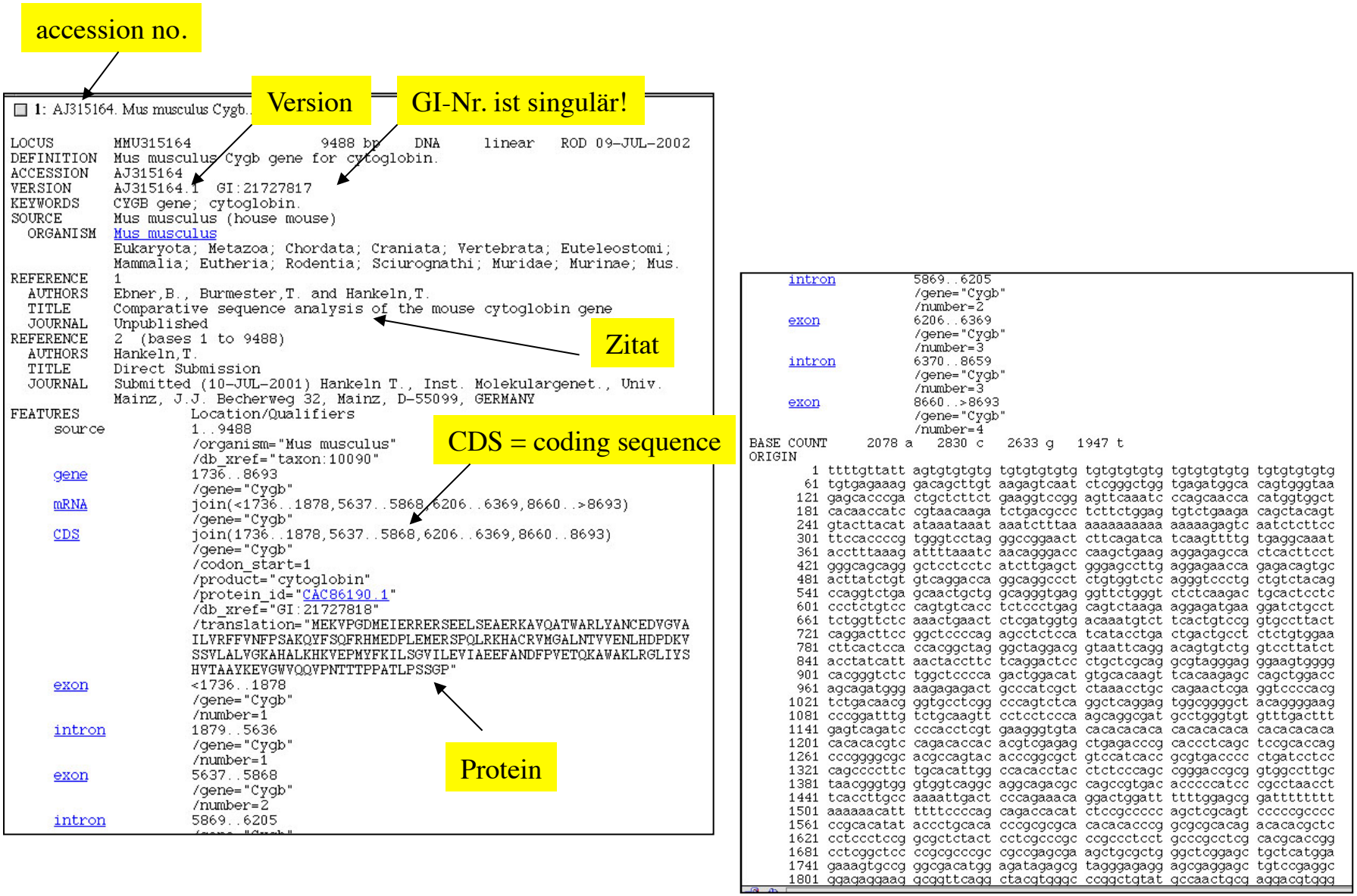
sowie ALU (repeats), VECTOR, YEAST, MITO, PAT(ente) und mehr

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

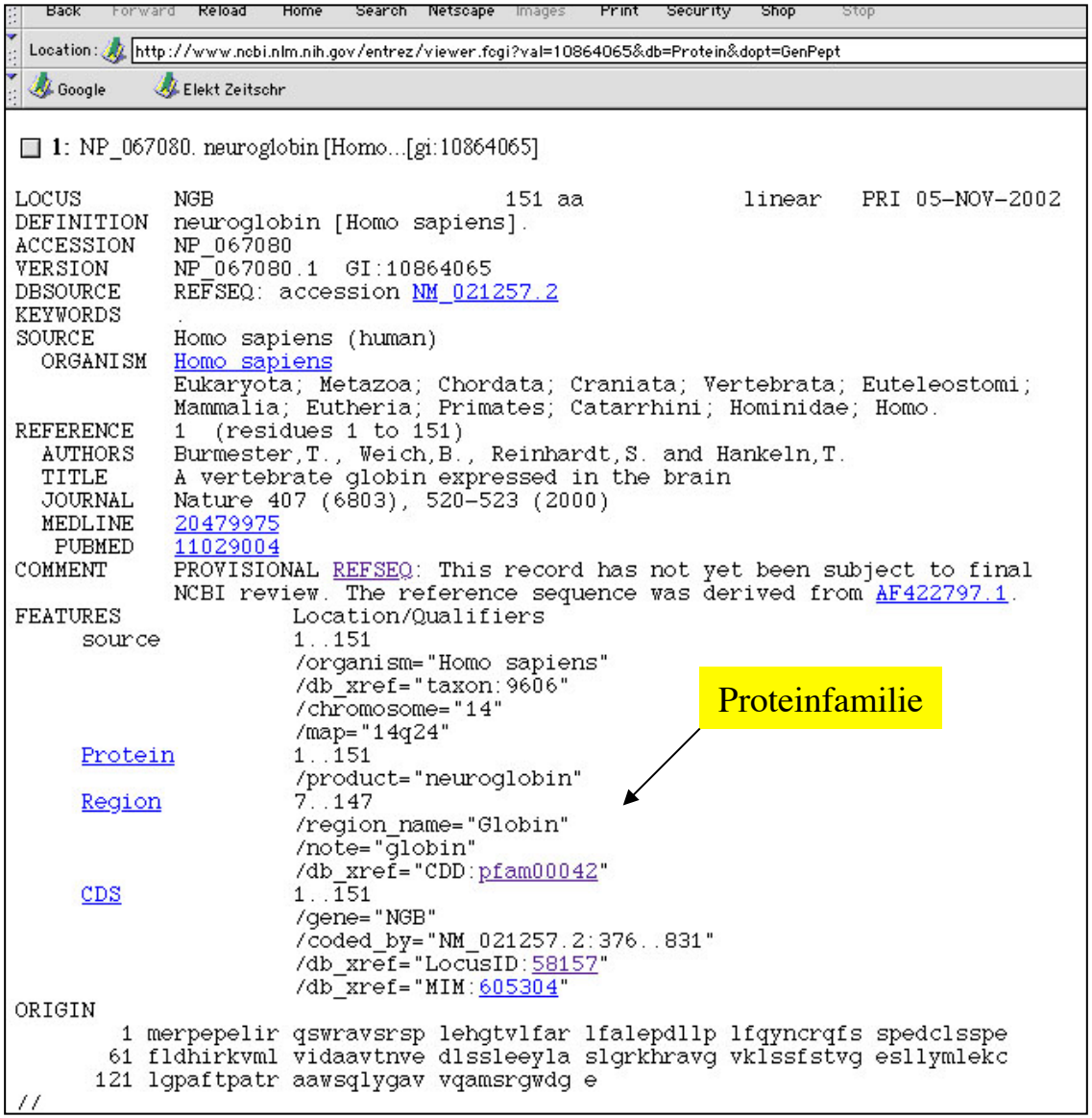
Division	Description	Release 191 (8/2012)	Annual increase (%)[a]
Taxnomic divisons			
SYN	Synthetic	928 200 038	494.2%
PHG	Phages	84 079 451	34.4%
ENV	Environmental samples	3 374 433 548	32.1%
VRL	Viruses	1 429 464 786	21.1%
BCT	Bacteria	8 439 854 434	21.0%
PLN	Plants	5 481 470 133	15.6%
MAM	Other mammals	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%

[a]Measured relative to Release 185 (8/2011).

Ein GenBank-Flat File...



...die dazugehörige Proteinsequenz



...und das Ganze als „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

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

GCACACAAAACCAGTTTATTTCTCATGAATTTATAGGACCACAGTCAGCACAAAGCAAAT
GGTACAAGTGCAAATGGCTGGGGTGAGGGCGGGGGCCTCTGCTGCTGCCTTCTTTTTTTT
CCAGCTGCCTGAGTACTGGCACACAGCGGGCAGTGAATCTGAGGAGCTGTGGCCTCACAG
TCGCTGCAGGCTGAGCCAAAAAGGAGGGCAGGG

Quality: High quality sequence stops at base: 90

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

COMMENTS

Gewebe

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

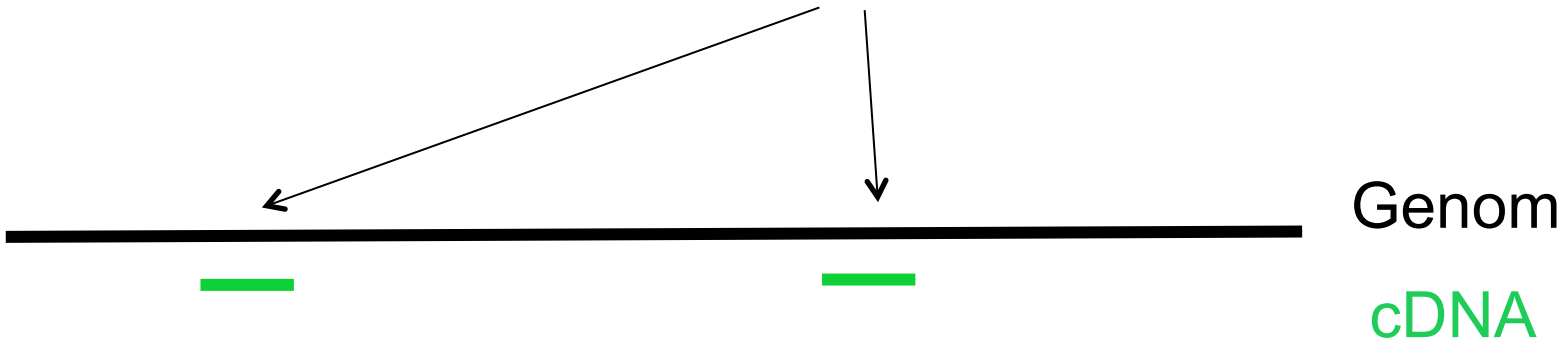
LIBRARY
Lib Name: Soares_total_fetus_Nb2HF8_9w
Organism: Homo sapiens
Develop. stage: 8-9 weeks
Lab host: DH10B
Vector: pT7T3D-Pac (Pharmacia) with a modified polylinker
R. Site 1: Not I
R. Site 2: Eco RI
Description: 1st strand cDNA was prepared from mRNA obtained from pooled 8-9 week (total) fetus material with a Not I - oligo(dT) primer [5'
TGTTACCAATCTGAAGTGGGAGCGGCCGCTTAATTTTTTTTTTTTTTTTT 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!!

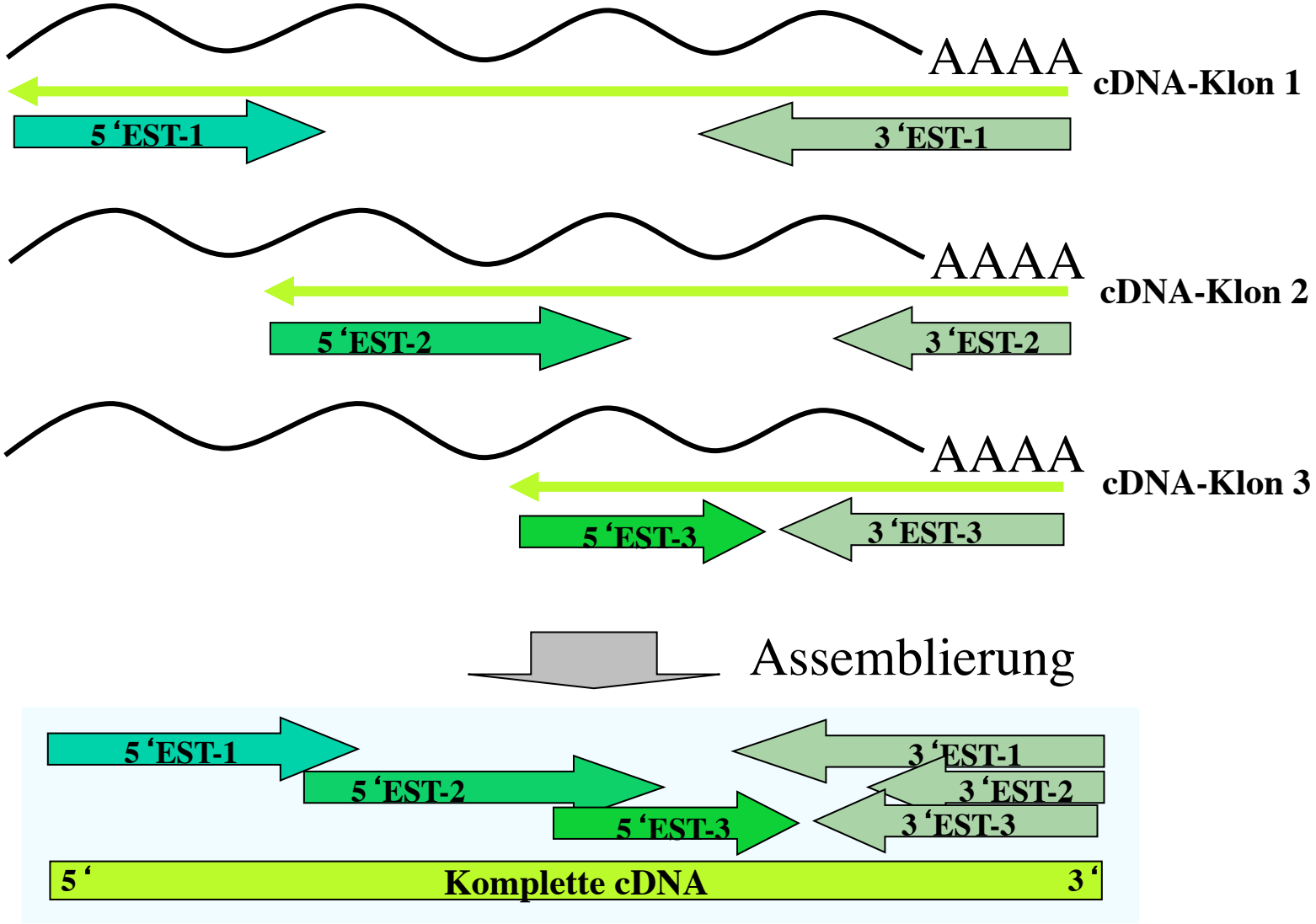
- ESTs enthalten viele Lesefehler!

- nicht komplett gespleißte cDNAs

- genomische Verunreinigung
  und Artefakte des Klonierungs-
  vorgangs (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, Genomseqenz, 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

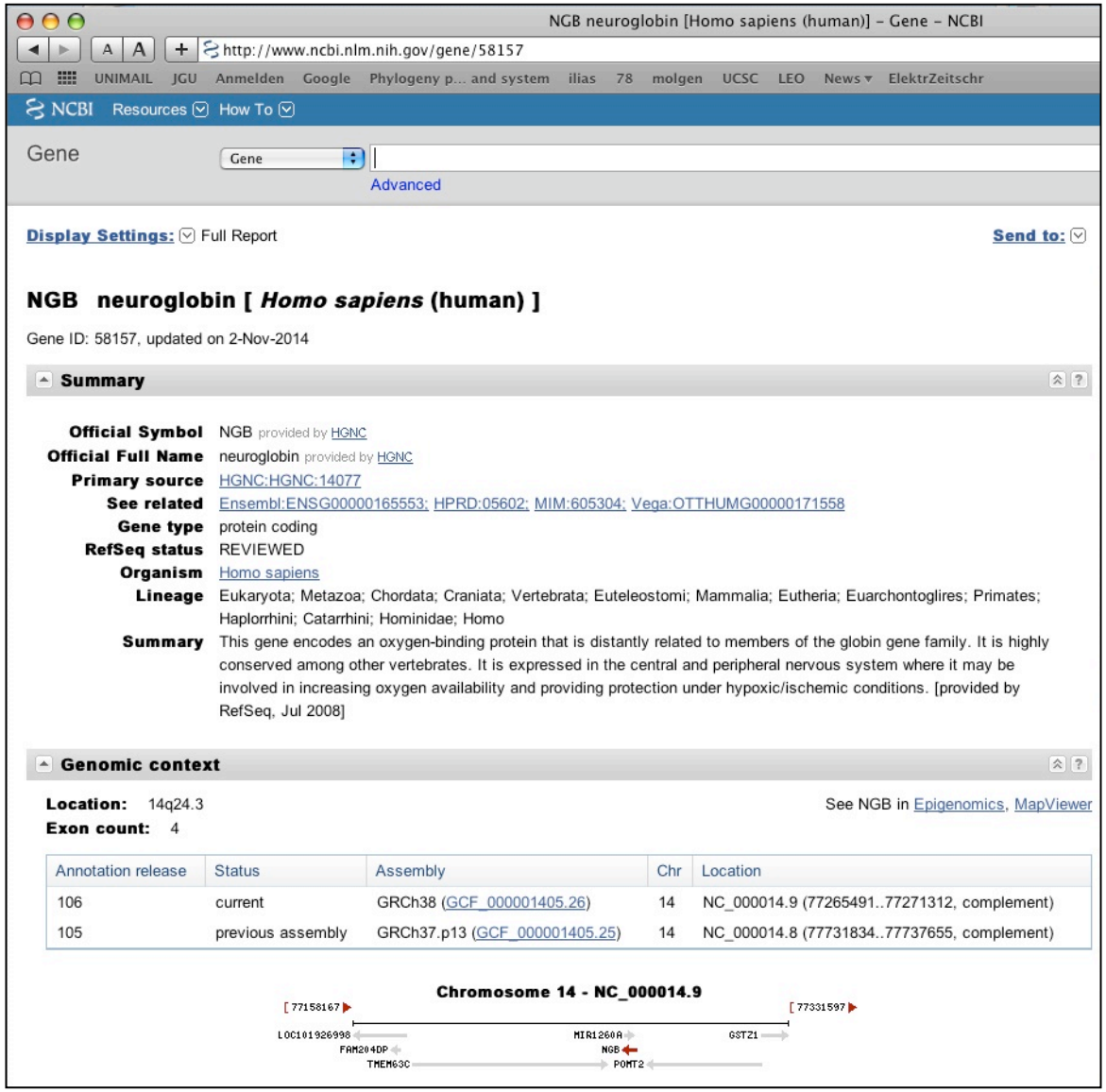
www.ncbi.nlm.nih.gov/RefSeq/

RefSeq-
„die Referenz unseres Wissensstandes“

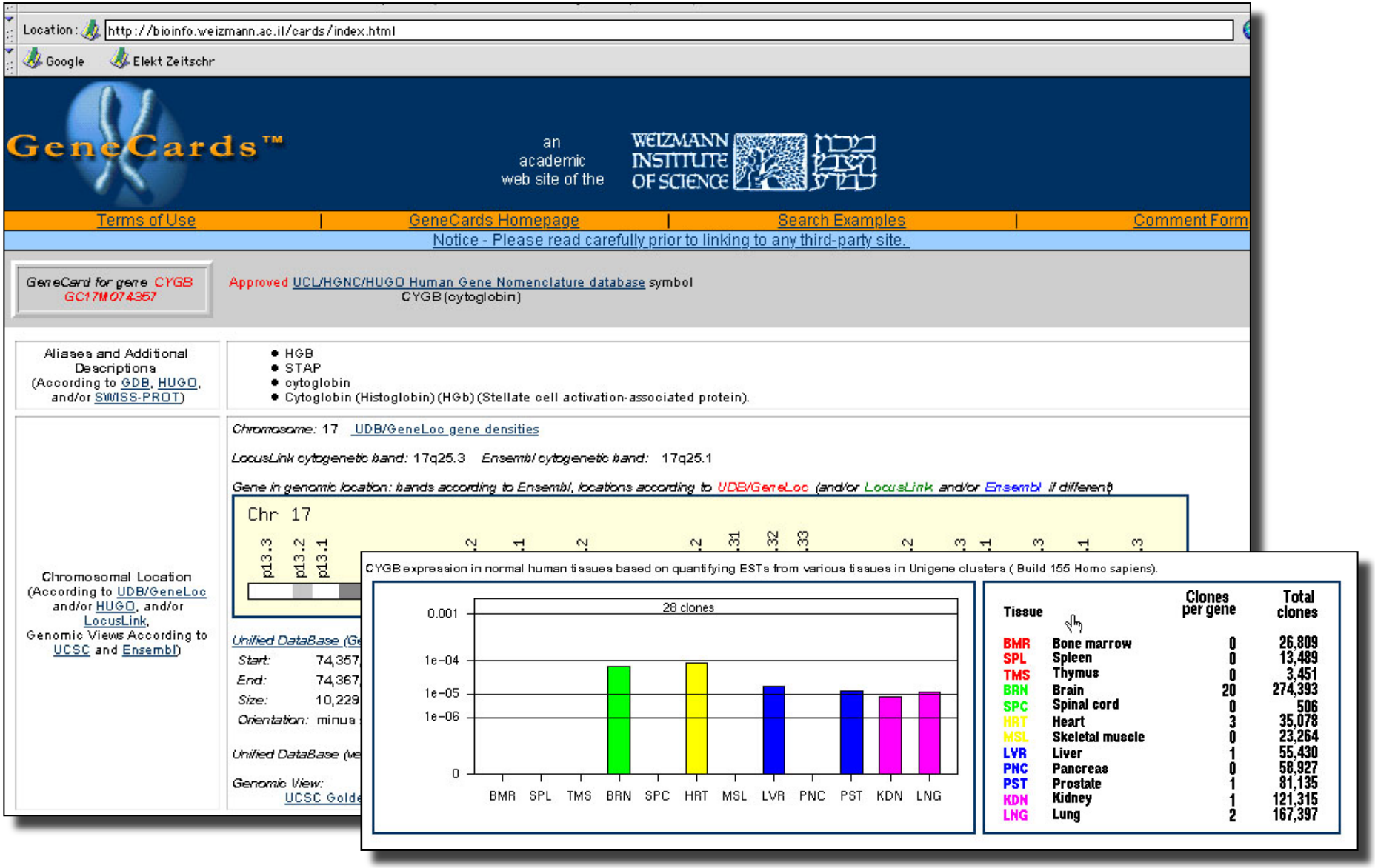
Key Characteristics of GenBank *versus* RefSeq

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“



GeneCards = Alternative zu EntrezGene

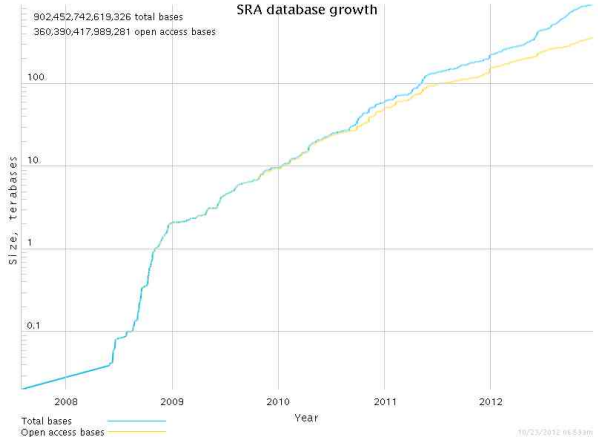


Für Entdecker!

Die Sequenzen hier sind meist unpubliziert…

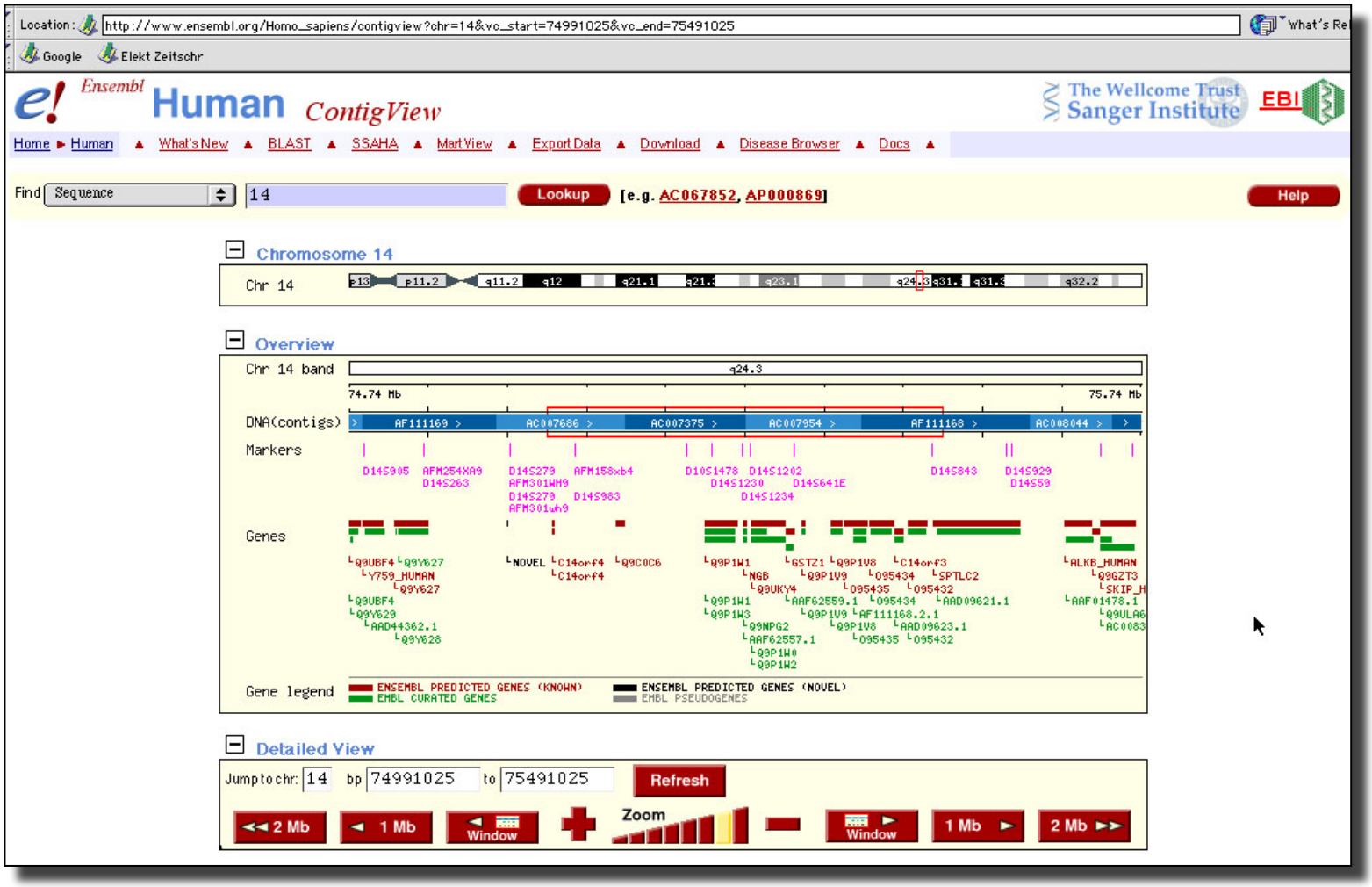
- Trace Archive: unannotierte Sanger-Reads aus EST-und Gesamtgenom-Projekten (>500 Spezies, 2.1 Milliarden Reads)

- Sequence Read Archive: NGS reads diverser Technologien (Tbytes!)



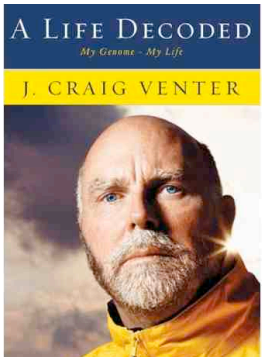
Spezielle annotierte Datenbanken für Genome

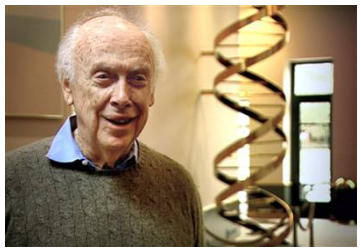
www.ensembl.org

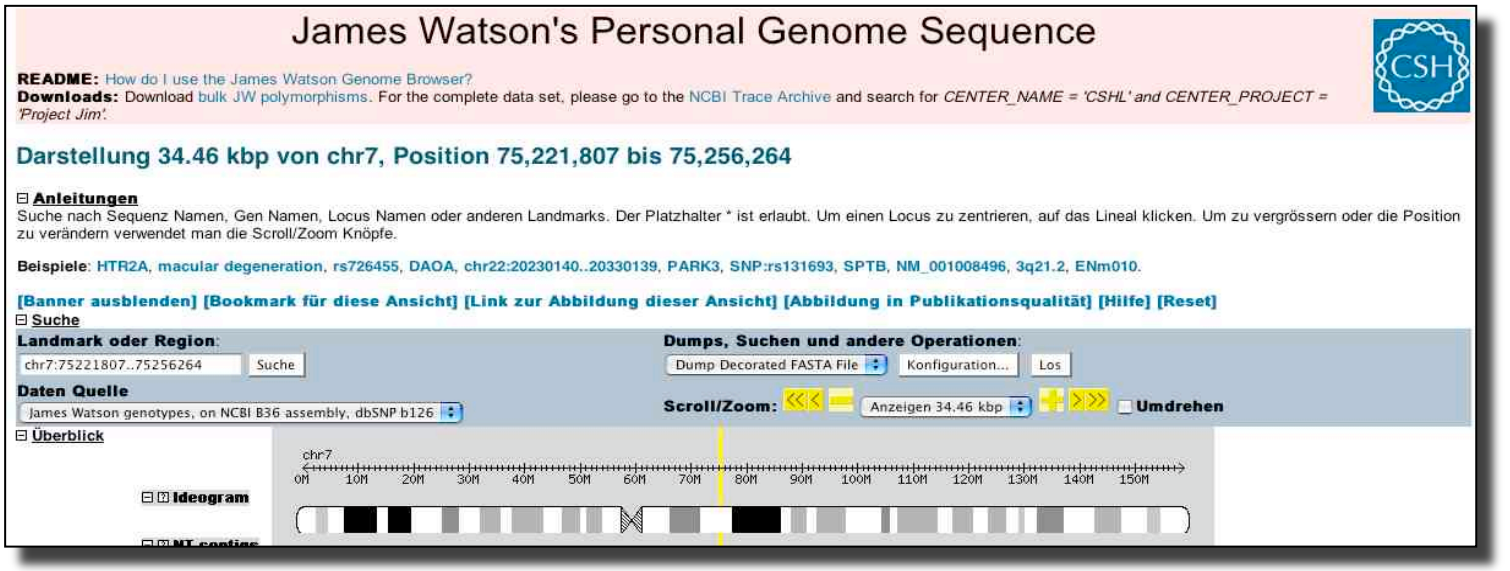


Personal Genomes

http://huref.jcvi.org/







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/

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

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